IMAGE ANNOTATION AND TOPIC EXTRACTION USING SUPER-WORD LATENT DIRICHLET ALLOCATION

DISSERTATION

George E. Noel III, Lieutenant Colonel, USAF

AFIT-ENG-DS-13-S-03

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-DS-13-S-03

IMAGE ANNOTATION AND TOPIC EXTRACTION USING SUPER-WORD LATENT DIRICHLET ALLOCATION

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

George E. Noel III, B.S.C.S., M.S.I.R.M.

Lieutenant Colonel, USAF

September 2013

AFIT-ENG-DS-13-S-03

IMAGE ANNOTATION AND TOPIC EXTRACTION USING SUPER-WORD LATENT DIRICHLET ALLOCATION

George E. Noel III, B.S.C.S., M.S.I.R.M.
Lieutenant Colonel, USAF

Approved:

Dr. Gilbert L. Peterson (Chairman)	Date
Dr. Christoph C. Borel-Donohue (Member)	Date
Dr. Jeffrey D. Clark (Member)	Date

Accepted:

Dr. Heidi Ries Interim Dean, Graduate School of Engineering and Management	Date

AFIT-ENG-DS-13-S-03

Abstract

The rapid expansion of data size and adoption of digital devices makes finding relevant digital forensics artifacts challenging. Automated processes frequently result in excessive false positives that strain human analyst resources. Additionally, these automated processes are typically limited to a specific data domain, such as text or images. Cross-domain solutions improve information retrieval, though work in these areas remains sparse. This research presents a multi-domain solution that uses text and images to iteratively improve automated information extraction.

The automated image annotation and latent topic extraction model operates in three stages. Stage I uses local text surrounding an embedded image to provide clues that help rank-order possible image annotations. The annotations have been demonstrated to outperform Automated Linguistic Indexing of Pictures in Real Time (ALIPR), one of the dominant generalized image annotation algorithms available. These annotations are forwarded to Stage II, where the image annotations from Stage I are used as highly-relevant "super-words" to improve extraction of topics. Stage II demonstrates improved topic extraction over several other similar latent topic models. The model probabilities from the super-words in Stage II are forwarded to Stage III where they are used to refine the automated image annotations developed in Stage I. By rank-ordering the super-words using model probability, Stage III demonstrates that the cross-domain model improves image annotation while iterating the model improves results further.

Finally, this research applies these techniques to a set of real-user hard drives. We demonstrate that latent topic models offer some advantages over standard document query techniques using real-world noisy digital forensics data.

Table of Contents

Page

Page

Page

List of Figures

List of Tables

List of Acronyms

ALIPR Automated Linguistic Indexing of Pictures in Real Time 130

AMT Amazon Mechanical Turk 133

BMMLDA Blob Multimodal Latent Dirichlet Allocation 132

CBIR Content-Based Image Retrieval 12

CPT Conditional Probability Table 146

DCT Discrete Cosine Transform 17

EM Expectation-Maximization 30

EMD Earth Mover's Distance 131

GM-LDA Generalized Mixture Latent Dirichlet Allocation 36

GPS Global Positioning System 1

INEX INitiative for the Evaluation of XML Retrieval 131

LDA Latent Dirichlet Allocation 144

LDCC Latent Dirichlet Co-Clustering 34

LDSEG Latent Dirichlet SEGmentation 40

LSA Latent Semantic Analysis 132

MCMC Markov Chain Monte-Carlo 69

NMI Normalized Mutual Information 43

OCR Optical Character Recognition 51

pLSA Probabilistic Latent Semantic Analysis 144

RDC Real Data Corpus 133

SAM Spherical Admixture Model 40

SVD Singular Value Decomposition 29

SIFT Shape-Invariant Feature Transform 17

SIMPLIcity Semantics-sensitive Integrated Matching for Picture LIbraries 20

SWLDA Super-Word Latent Dirichlet Allocation 130

TF-IDF Term Frequency - Inverse Document Frequency 62

USPTO U.S. Patent and Trademark Office 44

WTLDA Weighted Term Latent Dirichlet Allocation 132

Acknowledgements

I would like to express my sincere appreciation to my research advisor, Dr. Gilbert Peterson, for his dedication, unerring guidance, and patience throughout this dissertation effort. This work could not have been possible without his faithful mentoring and wisdom. I would also like to thank Dr. Simson Garfinkel for providing data and assistance during the experimentation on the Real Data Corpus. Finally, I would like to thank my committee members, Dr. Christoph Borel-Donohue and LtCol Jeffrey Clark for their time, effort and advice during this process.

George E. Noel III

IMAGE ANNOTATION AND TOPIC EXTRACTION USING SUPER-WORD LATENT DIRICHLET ALLOCATION

I. Introduction

A number of recent high-profile digital forensics cases highlight the growing importance of digital data in criminal investigation. The corruption trial of Illinois Governor Rod Blagojevich utilized digital voice recordings to cement their case [24]. Former Arizona representative Rick Renzi was found guilty in a complicated land swap deal partially due to evidence extracted from his business documents [84]. Digital forensics is a critical element of online crime investigations [98]. Even more so, media exploitation plays a critical role in the Department of Defense mission. Often, rapid exploitation of digital media can produce time-sensitive intelligence that could mean the difference between life and death, as recently highlighted by the raid on Osama bin Laden's compound [83].

The challenge in any large data repository is akin to finding a needle in a haystack. Amidst the photos of a family, their dog, and documents describing their favorite sports team's statistics may lie the Global Positioning System (GPS) coordinates of a key meeting between leaders of a terrorist organization. While humans may be able to flag this information once seen, searching for and finding that information could take too long. Computers can sift through megabytes of information in mere seconds. Unfortunately, they lack the intuition necessary to understand complex concepts and identify applicable topics that exist beyond basic keyword searches.

Data mining research aims to improve on the computer's ability to extract semantically meaningful information from large document corpora. The main goal,

according to Frawley et al. [38], is to extract "implicit, previously unknown, and potentially useful information from data". For the forensics analyst, what information is considered useful may change with the situation or be difficult to define. Relevant results may take the form of text, still images, video, or audio in a number of different data forms. An effective tool must be capable of searching through a wide variety of formats, integrating information into topical clusters that can be leveraged by user searches. While audio and video is beyond the scope of this research, images and text are used to improve automated data search with a comprehensive search model.

Two areas of data mining, automated image annotation and text mining, have expanded rapidly in the last couple decades [22] [4]. Automated image annotation and Content-Based Image Retrieval (CBIR) have proven challenging, with one researcher [93] even claiming "no actual progress" in the labeling of images. Decades of research have been directed towards CBIR and automated image annotation techniques [22] and while CBIR research demonstrates accurate results during controlled testing [22], it produces decreased performance on real world data [93]. Typically, both image annotation and CBIR tools must choose between approaches that either sacrifice precision for generalized applicability or specialize and fail outside of their limited domains [87]. This research focuses on a number of these problems and proposes a model that has annotation accuracy beyond the state of the art.

Text mining attempts to add structure to an unstructured document corpus. A number of techniques have been developed toward this goal with varied success [58], from classification techniques that make a discrete decision on categorical assignment to latent topic extraction. Topic-based queries are particularly applicable to this problem domain since they provide a robust method for isolating interesting topics amid numerous overlapping ones and decreasing data density of a returned query. Latent topic extraction techniques, such as Latent Dirichlet Allocation (LDA) [14],

have demonstrated high accuracy when compared against human topic assignment. Additionally, latent topic techniques handle topic changes within a document and provide flexibility for topic assignment at progressively smaller increments, such as paragraphs or sentences [109].

One solution may be to leverage information in multiple embedded media types to improve annotation and clustering. Research in multimodal fusion, or fusion of multiple data types, attempts to develop techniques for combining data domains, such as images and text, to extract refined conclusions. For example, video has demonstrated a strong link with audio in retrieval performance [53]. While multimodal fusion research is limited [22], it has helped in identifying scene changes [121] and story segmentation [139] in video. This was extended to text by Shafiei, et al. [109] to topically segment text documents by paragraphs and sentences. Others utilize a joint word-image modeling approach [78] [7] [60] [13] [70] that uses word and image region co-occurrence to link topically-significant entities.

Likewise, images provide valuable context for a document and text can help prompt image annotation. A growing body of research [7] [14] [78] uses surrounding text, image blobs, query logs and captions to link image blobs and text using latent topic models, such as LDA. This can be expanded to not only text, video, and images, but cell phone usage [32].

1.1 Research Hypothesis

This research tests the hypothesis that automated image annotation and latent topic extraction can each be leveraged to improve performance in the other. The research is accomplished using a multi-stage model to validate three main assumptions. First, local context surrounding an image can be used to improve automated image annotation performance over existing generalized annotators. Second, image anno-

tations can be used to improve latent topic extraction through the LDA algorithm. Third, posterior probabilities from the LDA model can be leveraged to improve image annotation and the model can be iterated as necessary to further improve results. Finally, this research hypothesizes that some techniques within this dissertation can be effectively used on a corpus of real-user data.

1.2 Challenges

Testing the hypothesis above requires addressing a multidisciplinary array of challenges. Automated image annotation is complex due to the unstructured nature of image features. Latent topic extraction has a tendency to focus on broader topics at the expense of smaller, potentially important topics. Fusing images and text into a single context requires mapping one domain to the other where, often, these maps are not obvious. This section briefly introduces some of the challenges this research will address.

1.2.1 Image Annotation.

The generalized image annotation methods [130] [140] [19] are designed to work across a broad spectrum of images, avoiding the a-priori assumptions about the data made by domain-specialized methods. Utilizing versatile feature sets ranging from basic color space to generalized frequency measures, these annotators use inherent differences in image features to discriminate between categories. Unfortunately, successful annotation requires high intra-category clustering with adequate inter-category separation. As the number of semantic tags increase, categorical separation becomes more difficult due to overlap.

Conversely, specialized classifiers often perform well within their domains, correctly classifying between indoor or outdoor images [117], city versus landscape [120],

or other specific domains. The drawback to specialized image annotation methods is they require a-priori assumptions about the data that, for a general image set may be incorrect. Demonstrating the performance of an algorithm then requires ensuring the a-priori prerequisites exist within the test data. One criticism [93] of image tagging research blames careful data selection on the disparity between the relatively good performance found in modern image tagging research and the poor performance when subjected to real world data. Müller, et al. [86] demonstrated how careful image and category selection within the Corel picture database can make a significant difference in performance of an image annotator. Yet without some level of manual data selection, real world data noise may obscure incremental gains in algorithmic performance.

1.2.2 Latent Topic Extraction.

Active research in extracting latent topics from text using a generative model, such as LDA [14], assumes all data points were generated from a model with hidden parameters. Unfortunately, predicting hidden model parameters is often an intractable problem, requiring approximation techniques that are negatively affected by outliers. Some models are sensitive to probabilistic islands [138] caused by high frequency spikes. Data smoothing techniques are often used to mitigate these issues [14] [57] [26], pruning outliers and smoothing spikes to produce a more accurate model [69]. Unfortunately, this destroys data that was, in concept, generated by the model.

There are certain cases where low frequency data points must be emphasized over the noise threshold based on their significance to the model. One example of this problem involves the melding of text mining and image annotation techniques. Image annotation tools use image features to tag images with words. Some image annotation tools [73] generate word tags for a query image based on a previously

trained model. A number of these tags may be irrelevant to the actual image based on the context, but a handful may be highly relevant. Assuming that an image tag carries more weight when compared to a single word drawn from the document, these tags should have a stronger influence over latent topic extraction. Merely inserting the image tag into the document text obscures this impact due to the low frequency of image tags. The challenge is to effectively raise the probabilistic influence of these image-generated words without disrupting latent topic extraction within the remaining words of the corpus. This could be accomplished through pruning, which risks obscuring the model's true parameters. Another technique involves emphasizing the influence of these significant words to minimize the impact of the less significant data on the model.

1.2.3 Multimodal Fusion.

Different data domains often lack direct mapping between similar entities in each domain. Automatically extracting these mappings can be a complex task prone to a high error rate. Image and text region co-occurrence provides one means for automated image annotation, but it implies a strong link between image blob frequency and word frequency. Research seeking to address this limitation uses word synonym relationships to refine image tags [8] and a common sense dictionary to link words [76]. Object recognition provides another potential, seeking to find specific objects within an image using a signature or template [100]. Ontology-based object recognition has demonstrated some success [106] [128] [3] at identifying general image regions such as sky, water, car parts, and mountains. Object recognition methods, however, tend to require complex training to account for different lighting, angles, and sizes. Blurry or noisy images may create a significant problem with false positives and false negatives. Additionally, object recognition requires a large, manually annotated image database

for each object category.

1.3 Research Outline and Part I

This research tests the hypothesis in two parts. In Part I, a model is defined that leverages information within text and images to iteratively improve both image annotation and latent topic extraction. Part II analyzes the Real Data Corpus (RDC), a corpus of real-user hard drives, to determine which techniques defined in this research could apply towards the digital forensics domain. Part I is accomplished in three stages, each described in the following sections.

1.3.1 Stage I: Automated Image Annotation.

Stage I tests the hypothesis that local context surrounding an image can be used to improve automated image annotation. This is accomplished by using the text surrounding an image to select an appropriate image annotation algorithm, then rank-orders possible annotations based on the results of those algorithms. First, words are converted into into synonymous sets (synsets), or word meanings, using WordNet [34] to identify all possible contextual meanings for each word. Those synsets with meanings related to people (e.g. doctor, philanthropist, etc) are tested using face detectors. All other synsets not related to people are tested against a set of specialized synset-based signatures generated from ImageNet [27] images and rank-ordered according to their estimated contextual accuracy. Since the purpose of this research is to develop methods that work on general noisy data, a wide range of Wikipedia articles [28] with embedded images is chosen for testing. Categories are selected based on their likelihood of including topically-oriented images, though no other data selection is performed beyond the automated pruning of sparse data.

The results are compared against the Automated Linguistic Indexing of Pictures

in Real Time (ALIPR) algorithm [73], a popular generalized image annotation algorithm. Results from Stage I demonstrate that leveraging WordNet and using local context, even with simple image annotation techniques, significantly improves results over the ALIPR algorithm.

1.3.2 Stage II: Super-Word Latent Dirichlet Allocation.

Stage II tests the hypothesis that image annotations can be used to improve automated topic extraction from a document corpus. The developed Super-Word Latent Dirichlet Allocation (SWLDA) model assumes a generative process based on LDA, where the model is assumed to have generated topics, documents, and words based on multinomial distributions with hidden parameters. In LDA, the priors for a multinomial distribution of words and super-words are generated via a Dirichlet distribution. The Dirichlet distribution has been demonstrated [14] to produce more reasonable priors for the multinomial distribution than other methods. By using independent Dirichlet priors for the separate word and super-word multinomial distributions, the super-words have greater influence on topic extraction. Unlike multimodal LDA models, a common vocabulary links the probabilistic influence of words and super-words. Results indicate statistically-significant improvements in document topic clustering when super-words are incorporated over standard LDA with text only.

Stage II is tested against an assortment of similar algorithms. Latent Semantic Analysis (LSA) [26] provides the baseline measure while LDA [14] is the primary comparison algorithm. Other algorithms include Blob Multimodal Latent Dirichlet Allocation (BMMLDA) and Weighted Term Latent Dirichlet Allocation (WTLDA), all described in further detail in Chapter V. Results indicate that SWLDA does as good or better than the test algorithm using document retrieval measures.

1.3.3 Stage III: Image Annotation Refinement.

Stage III tests the hypothesis that posterior probabilities from the model defined in Stage II can be used to further improve image annotations defined in Stage I. This process is iterated as many times as necessary to converge on a solution. Stage III results are compared against Stage I and II results to measure performance improvements. While Stage II results did not improve significantly, the Stage I annotations improved in average relevance.

1.4 Part II: Evaluation on Real-User Data

One of the primary uses for the techniques outlined in Part I is in searching for relevant documents within user data. Real-world user data in a digital forensics context is a challenging domain, often lacking structure or cohesiveness. Digital forensics is one area where the problems defined above are taking shape at an increasing pace. Part II analyzes the techniques within Part I from a digital forensics practitioner's point of view, using real-world user data.

A recent article by Garfinkel [41] states that the rapid growth of digital media indicates digital forensics is at the end of a Golden Age. Referred to as the "big data problem" [89], the expansion of data storage device size, digital devices, and cloud storage creates a data surge that can result in poor decision making, duplicated efforts, lost sales, and low productivity [33]. For the forensics or intelligence analyst, these can translate to lost opportunities, failure to present incriminating or exonerating evidence or, in extreme cases, even loss of life.

Preserving and preparing data for processing can be a daunting task. While a long and involved process [71], the bottleneck typically results from a lack of analyst time [88]. Unfortunately, today's analyst must deal with a variety of devices, from hard drives, digital cameras, thumb drives, cell phones, and tablets, among others. Cross-

drive analysis offers techniques to extract and correlate information across multiple data sources, potentially decreasing analyst burden.

The latent topic models described in this dissertation offer a potential solution towards effectively decreasing analyst time. By removing large numbers of files from the search space and organizing those files into topics, investigations may become easier. Latent topic models, such as the Latent Dirichlet Allocation [14], have been demonstrated successfully on large file corpus [127] [48] [94], though they have not been tested on real world storage devices. The RDC [42] offers a robust set of real-world disks from a variety of different countries and provides a valuable environment for testing topic models efficacy in the digital forensics domain.

Part II makes two primary contributions. First, it analyzes the RDC [42] from a digital forensics standpoint to identify areas for potential research and assesses how topic models could best be used. Second, it uses the RDC to compare traditional regular expression keyword search against the LDA. This is accomplished by first comparing document retrieval using regular expression keyword search against retrieval using LDA. Finally, automatic topic extraction using LDA is analyzed for accuracy.

1.5 Assumptions

Clustering of data in some form is required for accurate classification. This research assumes the data chosen includes latent topics that can be extracted using document and image features. It assumes the categories chosen by the INitiative for the Evaluation of XML Retrieval (INEX) 2007 researchers are accurate and that image annotations performed by the Amazon Mechanical Turk (AMT) workers accurately represents the opinions of the average human. The SWLDA model assumes a generative model exists with parameters that can be estimated, while the image

annotation methods assume patterns in features that provide adequate similarity measures, matching human semantic categories. Additionally, the image annotation model assumes the words surrounding an image are topically relevant to the image itself.

1.6 Document Layout

Chapter II presents a review of automated image annotation, text mining, and latent topic methods considered for this research. It provides a summary of comparison testing performed against several methods and justification for the techniques chosen. Chapter III provides a description of the overall system design, which describes how the different stages presented in the following three chapters fit together. Part II describes the experimental design, testing and results on the RDC in Chapter VII. Following this is a discussion of the overall research findings and conclusions.

II. Background

This research hypothesizes that automated image annotation and latent topic extraction can each be leveraged to improve the performance of the other. The three stages defined in Chapter I require a multidisciplinary approach that leverages specific advantages of the automated image annotation and latent topic extraction domains. The implied assumption is that we can take advantages of the strength of each approach while minimizing their limitations.

This chapter provides background on several areas essential to the presented data mining and retrieval approach. First, it is assumed most images will not include robust tagging information. Successful image annotation therefore requires analysis of image features, driving a discussion of Content-Based Image Retrieval (CBIR) and automated image annotation. Following this is a discussion on text mining and comparison techniques which provides a foundation for the Super-Word Latent Dirichlet Allocation (SWLDA) model discussed in Chapter V. Other applicable techniques are also discussed, including ensemble classifiers, multimodal latent topic extraction, among others, along with some of the complications with those algorithms. Finally, large corpus search is reviewed along with potential solutions for improving document search and retrieval accuracy.

2.1 Content-Based Image Retrieval / Automated Image Annotation

Many modern data repositories consist of more than just text information. Web pages, word processor documents, spreadsheets–all can include embedded images, videos, or tables that provide context and help define topical focus of a document. Performing latent topic extraction while ignoring the images within the document skips potentially significant data. Likewise, to attempt to perform image annotation

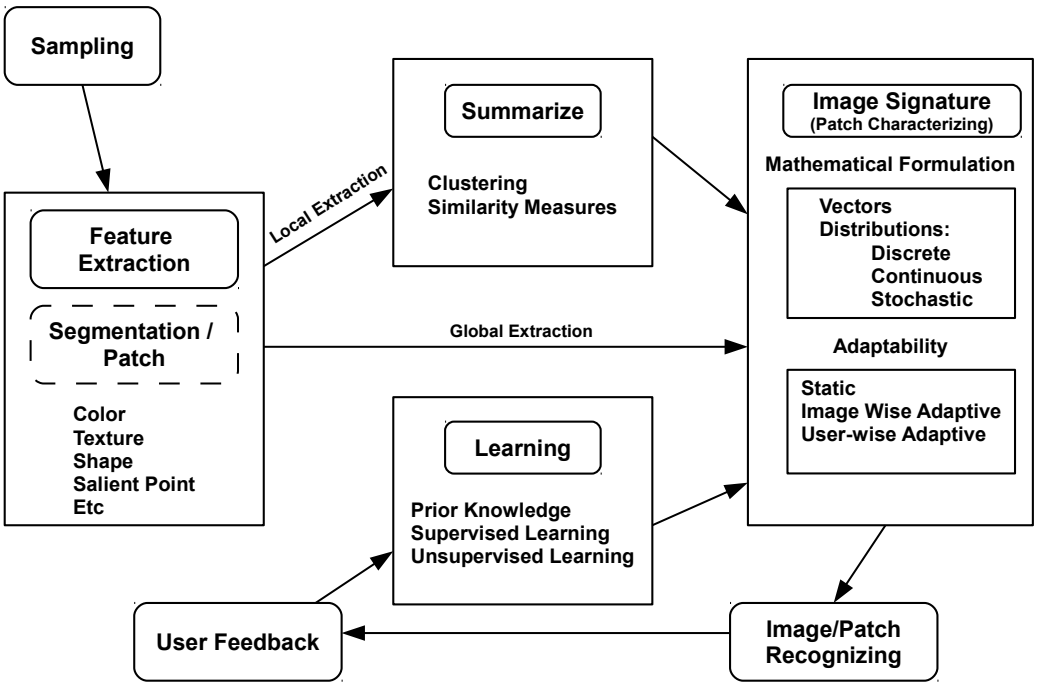


Figure 2.1: Content-Based Image Retrieval (CBIR) Taxonomy.

while ignoring local text abandons valuable context.

CBIR involves retrieving similar images based on image features derived from pixel information to respond to user queries. CBIR attempts to close the "semantic gap", or the difference between how users interpret an image and what can be derived by a computer using image features [112]. Automated image annotation is a similar approach with the goal of assigning metadata, such as keywords, to the image based on image features [22]. Users can then perform keyword searches or utilize combinations of approaches to retrieve relevant data.

Whether termed CBIR or automated image annotation, approaches tend to follow the taxonomy adopted and modified from Datta, et al. reproduced in Figure 2.1 [22]. Features are first extracted from the image and, in some cases, summarized to a representative set of data. A representative signature is developed based on the

image or set of images, then that signature is compared to others in an attempt to draw conclusions about unknown data.

Feature extraction is one of the most crucial steps in pattern classification since some features discriminate better than others. A wide variety of techniques exist and have been thoroughly covered in several surveys [22] [80]. Color space is a frequently used feature [73] [129] [22] based on proven ability to generalize and broad acceptance in the literature. Frequency is another common feature, using wavelets [129] [73] [130] or a variety of other means (see Section 2.1.1.2 for details). Shape, size, and salient point have also been used [22].

Adding these features potentially creates a high dimensional space. An image may consist of tens to hundreds of thousands of pixels. Large numbers of dimensions can impact algorithmic performance due to what is commonly called the "curse of dimensionality". Summarization techniques can be used to reduce the dimensionality to a reasonable level. Many utilize color space summarization with histogram bins [120] [117] or k-means averaging [73] [130]. Frequency may be measured using Daubechies or other wavelet transforms [129] [73] [130] and summarized similarly to color.

Summarizing the data in a way that maintains representative image information is not an easy task. The difficulty is compounded by the complexity of matching image features using a similarity measure. Many summarization techniques exist [22], from simple histograms to complex kernel-based signatures. Similarity measures for fixed-bin histograms are fast but imprecise [101]. Basic histogram comparison, such as bin-by-bin dissimilarity measures, provide a rough estimate of similarity, but fail to handle situations where histograms are off by one bin. Cross-bin similarity measures, such as Quadratic Form Distance, account for this, but still require fixed bins. The Earth Mover's Distance (EMD) [101] provides a signature-based approach that allows for inconsistently-sized histogram bins and measures the amount and

distance of "earth" that must be moved to convert one histogram to another. Too many or too few bins create performance issues, therefore, bin size must be fine-tuned to ensure performance. The next sections discuss available features and define those chosen for comparison testing.

2.1.1 Features.

The four most commonly-used features in CBIR include color, texture, shape, and salient point [22]. Color provides essential context that allows the viewer to distinguish between an image of a forest during the summer and one in the fall. Certain spectra are more prevalent in a natural scene versus urban images [120]. Texture is crucial for distinguishing between a red pie chart and a red rose–both may include similar color histograms but the rose would have a higher frequency than a solid color chart. Urban scenes include sudden changes in horizontal and vertical frequency with low frequency areas in-between, while a natural scene includes consistent moderate frequency in all directions. Shape is difficult to represent and many techniques are slow, though significant research has been accomplished with shape over the last two decades [22], including categorization based on object structures [21]. Salient points can be effective at representing local information, though it does not generalize well to large databases [107].

2.1.1.1 Color Space.

A number of color spaces exist, each solving a specific problem [15]. For instance, the NTSC color space was developed for television since it was easy to separate gray-scale information from color. This meant the same image could be displayed on both a color TV and black-and-white. RGB worked well for computer monitors that represented color as a combination of red, green, and blue pixels. While useful

for their particular purposes, they do a poor job of representing spectrum proximity as humans perceive it. HSV is far closer to human perception, using a mixture of tint, shade, and tone to describe a color, though it tends to bias towards painting representation [45] [62] [67]. The CIE-LUV and LAB color spaces use hue, saturation, and lightness measures, and come closest to approximating human vision. CIE-LUV is particularly useful when dealing with additive sources (such as a color monitor); CIE-LAB is best for reflexive surfaces [99] and has a solid body of research in CBIR and image annotation [101, 36, 75].

Transformation between color spaces can be accomplished using a variety of techniques. The RGB color space used in most image formats can be transformed to the CIE XYZ color space using a 3x3 matrix transform given in Equation 2.1. Conversion to CIE LAB space is then defined by Equation Set 2.2 where X_n, Y_n and Z_n are the tristimulus values for the constant energy white point, typically $X_n = Y_n = Z_n = 1/3$ [104].

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.1}$$

$$\begin{aligned} L &= \begin{cases} 116 * (Y/Y_n)^{1/3} & \text{where } Y/Y_n > 0.008856 \\ 903.3 * (Y/Y_n) & \text{otherwise} \end{cases} \\ a &= 500 * (f(X/X_n) - f(Y/Y_n)) \\ b &= 200 * (f(Y/Y_n) - f(Z/Z_n)) \\ f(t) &= \begin{cases} t^{1/3} & \text{where } t > 0.008856 \\ 7.787 * t + 16/116 & \text{otherwise} \end{cases} \end{aligned} \tag{2.2}$$

2.1.1.2 Frequency Measures.

Frequency is essential for distinguishing between a red car and a red pie chart, given the same average color. Yet, defining frequency in a meaningful way is not straightforward. Initial attempts to describe texture used rough texture measures such as the 28 descriptors based on a gray-tone spatial dependence matrix [50]. These are relevant only for an entire region. Later research by Tamara, et al. [118] attempted to improve on this, defining six basic texture characteristics. These include coarseness, contrast, directionality, line-likeness, regularity, and roughness. Since then, a number of methods have been in use in an attempt to accurately quantize frequency information. Research using the Shape-Invariant Feature Transform (SIFT) [127] [136] [132] [78] rely on a 128 byte descriptor that summarizes the edge information in a region. Other research [136] [36] utilizes the Gabor feature set, calculating orientation and scale information at each pixel and consolidating these into a region or global feature vector. Still others utilize the Discrete Cosine Transform (DCT) [36] [117], a Fourier-related transform that can be used to measure a region-based frequency.

The Daubechies-4 fast wavelet transform represents localized changes in color better when compared to simpler wavelets (Haar) [23] and has proven performance in other CBIR research [129] [74] [130] [20] [61]. The Daubechies-4 wavelet utilizes two vanishing moments to represent the image frequency and, while initial experimentation indicates this works well, further testing is required to determine if more vanishing moments would represent the signal better (more vanishing moments work better for complex signals).

The fast wavelet transform utilizes two filters: a lowpass decomposition and a highpass decomposition. Depending on how they are used, frequencies can be extracted in the horizontal, vertical, or diagonal directions. In Equations 2.3 and 2.4 $\varphi(x)$ is the scaling function and $\psi(x)$ is the wavelet function. Variable h is considered

(a) Original Image. (b) Daubechies-4 Fast Wavelet Transform.

Figure 2.2: Image Frequency Example Using the Daubechies-4 Fast Wavelet Transform.

the father wavelet, while $h_\psi(n)$ is the highpass decomposition filter and $h_\varphi(n)$ the lowpass.

$$\varphi(x) = \sum_n h_\varphi(n)\sqrt{2}\varphi(2x - n) \tag{2.3}$$

$$\psi(x) = \sum_n h_\psi(n)\sqrt{2}\varphi(2x - n) \tag{2.4}$$

The various horizontal, vertical, and diagonal frequency bands can be calculated via Equations 2.5, 2.6, and 2.7 respectively [45].

$$\psi_H(x, y) = \psi(x)\varphi(y) \tag{2.5}$$

$$\psi_V(x, y) = \varphi(x)\psi(y) \tag{2.6}$$

$$\psi_D(x, y) = \psi(x)\psi(y) \tag{2.7}$$

Since frequency resolution is only gained by sacrificing spatial resolution and vice-

versa, each iteration of the filter produces $\frac{n}{4}$ coefficients. This is illustrated in Figures 2.2(a) and 2.2(b). The image of the lighthouse is processed using the Daubechies fast wavelet transform and results in the horizontal frequencies in the upper right, the vertical frequencies in the lower left, and the diagonal frequencies in the lower right. The upper left square visualizes the second application of the filter set. The three squares are one-fourth the size as they must sacrifice spatial resolution for frequency resolution in the lower frequencies. This is repeated a third and fourth application with the remaining image coefficients shown in the final upper left corner.

2.1.2 Summarization and Clustering.

The color and texture for each pixel can be represented by a vector of feature values. Utilizing data from each pixel would create an image vector with high dimensionality. The "curse of dimensionality" creates a sparse data set that requires an exponentially large data set for many methods to function properly. This limitation can be mitigated by reducing, or summarizing, the data to a set of representative vectors. Earlier examples [116] [115] addressed dimensionality issues with color histogram indexing. More recent methods reduce data to a representative mean vector. Unsupervised clustering techniques, such as k-means [52], organize vector means based on similar colors or frequencies. This, unfortunately, ignores pixel location or contiguous regions.

Segmentation by region maintains pixel neighbor information, attempting to group pixels into regions based on proximity and pixel value thresholding. The goal is a set of related objects or entities within the image such as boats and water. One of the earliest segmentation approaches [81] uses pixel vector data consisting of color and frequency to compute gradient flow for edge detection. A similar approach, Blobworld [17], combines color, texture, and positional information for each pixel, clustering

with an Expectation-Maximization algorithm to identify regions. Shi and Malik [111] recently defined a graph-based segmentation that treats each pixel as a node and color change as an edge. It attempts to segment the image using eigenvalues to determine the optimal cut configuration. Felzenszwalb and Huttenlocher [35] developed Efficient Graph-Based Segmentation, a faster method using a greedy approach, combining nodes based on edge thresholds. The threshold is based on a user set parameter divided by the total size of the current node. The larger a node becomes, the lower an edge must be before it no longer combines nodes.

Ideally, the regions maintain image entities more completely than pixel value clustering. Chen and Wang [19] were able to utilize a region-based fuzzy feature matching approach to conduct effective CBIR. Li, et al. [74] created a region-based similarity comparison that attempts to match sets of individual regions that eventually was incorporated into the Semantics-sensitive Integrated Matching for Picture LIbraries (SIMPLIcity) [130] image retrieval system.

Figure 2.3: Efficient Graph-Based Segmentation with Mean Threshold [35] [90].

This work uses the Efficient Graph-Based Segmentation [35] to produce image regions whose vector means represent features common to all images of a particular

synset. Efficient Graph-Based Segmentation utilizes a ratio of region size to threshold node merging based on edge gradient. Unfortunately, this requires parameter tuning in relationship to image sizes. In an image with numerous pixels, the threshold shrinks as pixel size increases, resulting in a high number of regions. An example of the Efficient Graph-Based Segmentation with Mean Thresholding can be found in Figure 2.3. This is a complex image resulting in a large number of regions.

Algorithm 1 takes in an image, along with parameters defining the functioning of the algorithm. The paramter min_size defines the smallest number of pixels that can be used to define a region. c defines a threshold of pixel differences that indicate when regions should be combined together. Parameter $mean\_dist$ defines the distance in the mean pixel value that two regions can be apart before they are combined together. This parameter handles large regions of very similar features that, because of their size, did not combine together as they should have. Finally, sd_factor indicates the minimum differences in standard deviation that two images can have before they are no longer combined together. This can be used to prevent regions with wide color variance from combining with regions having little variance just because their means are close.

The algorithm starts by creating a node for every pixel with edges based on adjacent pixels. Edge weights are based on pixel differences. Nodes are combined using a threshold that decreases as the region gets larger, so eventually large regions require very similar colors for a region to be combined. Further information on this algorithm can be found in [35].

An image or set of images can then be represented by a signature containing the vector information. Signatures can then be compared against other signatures to categorize images using discriminant methods, unsupervised clustering, supervised clustering, or a number of other methods.

Algorithm 1 Efficient Graph-Based Segmentation

Input: Image, min_size, c, mean_dist, sd_factor
Output: Pixel regions
Create graph $G = (V, E)$ where each pixel is a node V and adjacent pixel is an edge E.
Initialize edge weights to Euclidean pixel value distance
Sort E by non-decreasing edge weight
for all edges $e_i \in E$ with connected nodes n_i and n_j **do**
  if $weight(e_i) <= c/size(n_i)$ **and** $weight(e_i) <= c/size(n_j)$ **then**
    Combine nodes n_j and n_k connected by edge e_i into new node n_l
  end if
end for
for all nodes n_i in N **do**
  if $size(n_i) < min\_size$ **then**
    Combine node n_i with neighbor node of lowest threshold
  end if
end for
for all edges $e_i \in E$ with connected nodes n_i and n_j **do**
  Calculate μ and σ of pixels in the region represented by nodes n_i and n_j
  Calculate distance factor $d = mean\_dist * (sd\_factor/average(\sigma_{n_i}, \sigma_{n_j})$
  Calculate Euclidean distance between means $dist = Euclidean(\mu_{n_i}, \mu_{n_j})$
  if $(dist < d)$ **then**
    Combine nodes n_j and n_k connected by edge e_i into new node n_l
  end if
end for

2.1.3 Signature Comparison and Similarity Measures.

Once a representative signature has been built, a method must be identified to measure similarity. Much of the earliest research in CBIR and automated image annotation relied on distance-based functions for image comparison. Pixel values, frequency information, and position, among other features, can all be stored in vector form. Euclidean and Manhattan distance provide basic measures of similarity that can be used, along with other measures, to compare pixel information from one image to another [43]. Nearest neighbor has been used successfully in a number of papers [132] [44] [80] for its simplicity and discriminative ability. k-means and Expectation-Maximization of Gaussian functions has also demonstrated success in clustering images and image regions [31]. Yet, every spatially-oriented measure assumes relevance in distance calculations between data points. The challenge is to define a space where distance has meaning, which may be simple when using color as a global descriptor, but becomes challenging when using ellipses, cylinders, and spatial relationships [21]. Some of the more promising histogram or signature-based distance measures include Earth Mover's Distance (EMD) [101], Perceptually-Modified Hausdorff Distance [92], and the Signature Quadratic Form Distance [11]. We review EMD in more detail, due to its wide use in existing research [47] [119] and useful attributes for CBIR [101].

2.1.3.1 Earth Mover's Distance (EMD).

EMD [101] provides a signature-based approach that allows for inconsistently-sized histogram bins and measures the amount and distance of "earth" that must be moved to convert one histogram to another. It has been successfully used in CBIR to match local invariant features [47] and [119]. Where previous histogram comparison methods required a fixed bin count and width, EMD can handle adaptive-bin histograms, or

signature-based models. It does so by solving a transportation problem to determine the near-optimal amount and distance of "earth" required to convert one histogram to another. The transportation problem itself is intractable, requiring an approximation technique with relatively high algorithmic complexity. The OpenCV library [1] has an effective implementation of the function that is used in this research.

2.2 Ensemble Classifiers

An ensemble classifier is a set of classifiers whose individual decisions are combined in some way [29]. In many cases, the ensemble of classifiers are more accurate than individual classifiers, as long as the individual classifiers are diverse and produce better-than-random results [49]. In concept, an ensemble classifier allows for a complex modeling of the search space not possible with other classifiers such as k-means or discriminant analysis. It has been used in CBIR to select a small set of features among a wide range of possible features [113] [62], and to improve classification using domain knowledge [77]. In the image search space, there are sets of images that lend themselves to other classifiers better than others. The image classifier being proposed in Chapter IV depends on representative images that define a word. When the word is people-related, such as 'doctor' or 'nurse', it becomes difficult to develop a representative sample for each word that is adequately distinct from the other. This requires specialized algorithms for classifying images with humans as opposed to images without. Additionally, graphs or clip-art also will rarely look like images of their real-world counterpart and require optical character recognition or similar means to extract meaningful information. Chawla, et al. [18] found that intelligent partitioning of the data set can yield better results. By partitioning the data into disjoint sets, then normalizing the results at the end, each classifier can be optimized towards its particular purpose.

2.2.1 Graph and Clip Art Detection.

For classifiers largely based on patterns within colors or frequency, clip art or graphs pose problems. A clip-art image of a cow is going to have a very homogeneous color histogram with sharp spikes in frequency followed by very low frequencies. Contrast this with an image of a real cow which includes a variety of shades and fairly smooth textures. Intuitively, it follows that clip art and graphs must be treated differently than regular images. Popescu, et al. [95] proposed a simple method for identifying clip art and graphs using histogram analysis. The SIMPLICity system [130] uses sector analysis of high frequencies to detect graphs, or low frequency images. Once a graph or clip-art is detected, optical character recognition could be run on the image.

2.2.2 People Detection.

A number of techniques exist for detecting people in still images and generally fall into two categories. The first involves spectral analysis of the image looking for skin-tones. Skin color occupies a very narrow band, becoming almost a single blob in HSV color spaces [66]. Simple spectral analysis can detect skin with high accuracy regardless of lighting level or luminescence value [122]. The simplest method uses a straightforward threshold, however, Gaussian fitting has been successfully used by training with skin versus non-skin images [66].

The second involves identifying human features, such as faces and limbs. A Haar-like feature cascade classifier has been used successfully [123] to identify faces in images. Shape context has also demonstrated success at identifying people [108] using edges extracted with a Canny detector. Fleck, et al. [37] combined color skin detection with limb analysis to identify objectionable images based on body position. A wide variety of additional techniques [105] are available for people detection, ranging

from motion video techniques to still images.

2.3 Context-Based Image Annotation

Clustering image segments with text makes one large assumption–that segments of sky and water have topical relevance with other words in a document. Often, an image's topical relevance requires more than just information from a single segment. Linking segments into contextually-significant entities is challenging and it is often unlikely that documents with images containing clouds or sky are going to mention clouds or sky. It is likely the entire discussion is about the aircraft, tornado, or other entity that dominates the image. Hence, local context surrounding the image can provide crucial data to drive image annotation over simple corpus co-occurrence.

Most images are embedded in context information of some kind: within a named directory structure, linked in a web page, an attached to an e-mail or embedded within a word processing document. One of the first examples of utilizing context information in image tagging [134] uses Latent Semantic Analysis (LSA) to take advantage of image feature and word co-occurence. Barnard, et al. [7] tested several methods besides what has been covered, the first being an extension on Hofmann's hierarchical model for text [56] into multimodal representations. It uses soft clustering to arrange co-occuring text and image segments into a tree structure. Higher levels on the tree represent more general words, such as blobs representing the sky. Extending Latent Dirichlet Allocation (LDA) beyond the document corpus, Liu, et al. [78] link image blobs with text into a large-scale, parallel infrastructure designed for web image tagging. They utilize query logs to help facilitate effective image tagging, along with captions and surrounding text. All these techniques require a significant number of images and words to effectively extract latent topics [60]. This, in addition to the issues with unsupervised clustering of large data sets described in Section 2.6,

complicate effective utilization of these methods for small or large data sets.

Classifier selection is merely one piece; specialized classifiers must be able to leverage this contextual information to improve classification. Jin, et al. [63] use semantic knowledge about words and word relationships to refine annotation. Barnard, et al. [8] takes this a step further, using Wordnet synonymous relationships to group image tags. Lieberman and Liu [76] use semantic analysis and a common sense dictionary to identify sentence subjects and link them with words that commonly appear together.

Object recognition provides another potential, seeking to find specific objects within an image using a signature or template [100]. Ontology-based object recognition has demonstrated some success [106] at identifying general image regions such as sky, water, and mountains. Wang, et al. [128] utilized a manually-crafted ontological hierarchy for animal characteristics (texture, color, shape, etc). It was able to significantly improve animal detection in images over purely textual searches. Agarwal et al. [3] focuses on specific elements of an object, such as car wheels or the front grill. They automatically construct a vocabulary of vehicle parts and train a classifier to identify existence of these parts in unknown images. Chai, et al. [21] used edge detection to search for ellipses and quadrangles, identifying and discriminating between both cars and bicycles. Object recognition methods tend to require complex training to account for different lighting, angles, and sizes. Blurry or noisy images may create a significant problem with false positives and false negatives. Additionally, object recognition requires a large, manually annotated image database for each object category.

2.4 Text Mining

Where CBIR and image annotation must draw inference from ambiguous pixel information, text mining can take advantage of a known set of dictionary words.

Still, ambiguity complicates data retrieval. A simple keyword search text query often returns far more documents than are relevant, yet misses synonyms with higher relevancy. Document clustering attempts to isolate natural groupings that can be leveraged based on either a new unknown sample or user query. Clustering has been divided between discriminative and generative types. Discriminative clustering uses pairwise distances between data points and optimizes an objective function to produce optimal clustering. Contrast this with generative models that assumes an underlying generative distribution for the data and attempts to predict distribution parameters (maximizing fit). While many generative models exist, several have had significant success in recent years, including the Probabilistic Latent Semantic Indexing [57] and, more recently, Latent Dirichlet Allocation [14]. These utilize the "bag-of-words" assumption, disregarding word order within documents and document order within a corpus.

2.4.1 Hierarchical Clusters.

The challenge for clustering algorithms is to extract complex topics from a large, potentially diverse document corpus. Hierarchical clustering has been the dominant clustering technique [5] in text mining. It gradually clusters documents, or the smallest hierarchy, into a progressively larger hierarchy. While useful since the corpus can then be browsed, it does not scale well and performs poorly in a large document corpus [5]. Partitioning algorithms, such as k-means, are generally accepted to work better than hierarchical, especially when it comes to "bisecting" k-means [114]. Further refinement of partitioning algorithms, including Expectation-maximization, are also available. Graph-based algorithms have demonstrated some success, but many algorithms suffer from the scalability problem of hierarchical methods. While a number of other clustering algorithms have been applied to document clustering, an exhaustive

review is outside the scope of this dissertation.

Two clustering algorithms have demonstrated repeated success at clustering large document corpus. The Self-Organizing Map provides an efficient way to create an "internal representation" of the input signals and clustering data around a trained lattice of nodes [68]. It has the advantage of being relatively fast, can operate on large data spaces with high dimensionality, and is not as hampered as other methods by sparse matrices. LSA [26] utilizes Singular Value Decomposition (SVD) to take advantage of word co-occurrence. Probabilistic Latent Semantic Analysis (pLSA) [57] provides a statistical foundation for LSA by using a multinomial distribution with priors generated by a uniform distribution. LDA takes this a step further by using a Dirichlet distribution to generate priors and has been widely cited due to its simplicity and applicability across a wide range of data mining applications [14].

2.4.2 Latent Semantic Indexing.

Latent Semantic Analysis (LSA) [26] uses SVD to rotate the vector space matrix consisting of word frequencies in documents. SVD produces the joint probability model $P = \hat{H}\hat{\Sigma}\hat{V}^t$. In Blei, et al. [14], $\hat{U}$ corresponds to the generative probability of a document given a topic and $\hat{V}$ the probability of a word given a topic. Subsequently, the problem can be simplified by thresholding the lowest K singular values in Σ to isolate interesting topical clusters. pLSA [57] expands on LSA using a different objective function based on the likelihood function of multinomial sampling and attempts to maximize the model's predictive power. While effective, this algorithm provides a generative model for words but is unable to predict document generation. Additionally, the large, sparse matrix complicates necessary matrix inversion and increases algorithmic complexity.

2.4.3 Bayes Network.

A Bayes network is a probabilistic model typically represented as a directed acyclic graph that models a series of random variables and their relationships. Inferences can be made about nodes in the network given parent or child nodes, making the Bayes network a powerful tool and essential for the next discussion about LDA.

2.4.4 Latent Dirichlet Allocation.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model useful for extracting latent topics within a document corpus. It was originally defined by Blei, et al. [14] in 2003 and initiated a surge of research, including [48] [78] [109]. LDA models each item of a collection as a finite mixture over a latent set of topics. It is considered a generative model, meaning that documents and topics assumed to generate words according to the model's distributions. Unfortunately, the model parameters are usually not known; only the generated documents and words are available. Using estimation techniques such as Expectation-Maximization (EM) or Markov Chain Monte-Carlo (MCMC) simulations, these parameters can be approximated and used to predict future words or documents.

Like many of the other latent models, LDA makes the "bag-of-words" assumption, treating each document as a set of unordered words. Each document, utilizing the vector space model [103], can be represented as a vector histogram of word frequency. The method is illustrated by the plate diagram in Figure 2.4. The outer plate represents M documents while the inner plate represents the N words within the corpus. The prior probability Dirichlet parameter α and its generated multinomial distribution θ produce per-document topic distributions, while the multinomial distribution ϕ represents word topic distributions.

The generative model assumes the process defined in Algorithm 1 [14], where a

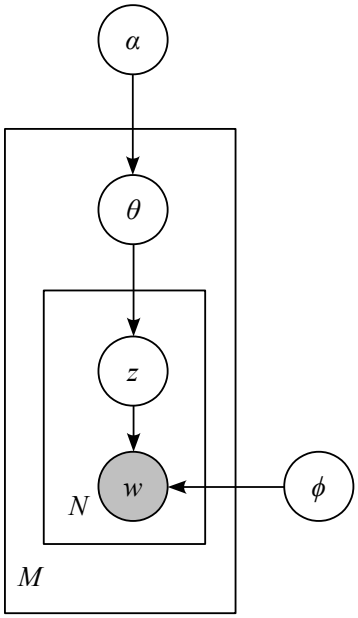


Figure 2.4: Latent Dirichlet Allocation Plate Diagram [14].

defined model generates documents, topics, and words according to the given probability distribution.

Algorithm 2 LDA Generative Process

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \phi)$, a multinomial probability conditioned on the topic z_n

$$q(\theta, z|, \gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{2.8}$$

The goal is to determine the parameters of the hidden generative model using Equation 2.8. In this equation, the Dirichlet parameter γ and multinomial parameters ϕ are the free variational parameters [14]. The matrix z includes word-document topic

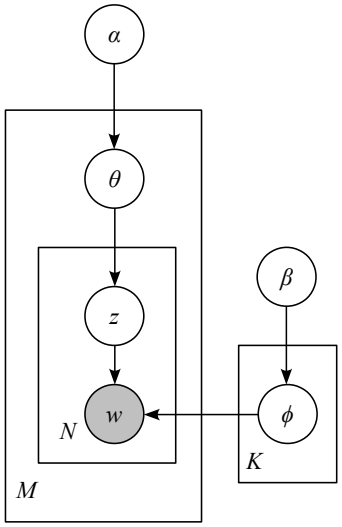


Figure 2.5: Latent Dirichlet Allocation with Smoothing Plate Diagram.

assignments and w the word list.

Most LDA implementations perform some form of data smoothing. This is demonstrated in Equation 2.5 by the additional Dirichlet distribution with priors β, representing the topic by vocabulary stochastic matrix. Data smoothing often improves results by eliminating outliers and smoothing frequency spikes but can result in loss of fidelity. This will be discussed further in Section 2.6.

Solving for the model parameters is an intractable problem [14]. However, several techniques have evolved to solve approximations of optimal parameter values. In the original LDA work [14], Blei, et al. define a Bayes approach where one seeks to define a lower bound for Equation 2.9. They do this using Expectation-Maximization to maximize the lower bound by adjusting the parameters α and β (Equation 2.10) during each iteration. This is, however, a relatively costly calculation when compared with Gibbs Sampling [48].

$$(\gamma^*, \phi^*) = \underset{(\gamma,\phi)}{\operatorname{argmin}} D(q(\theta, z|\gamma, \phi)\|p(\theta, z|w, \alpha, \beta)) \tag{2.9}$$

$$(\alpha, \beta) = \sum_{d=1}^{M} \log p(w_d|\alpha, \beta) \tag{2.10}$$

2.4.4.1 Gibbs Sampling.

Gibbs sampling offers a simple iterative approach to LDA parameter estimation that provides acceptable approximation. It uses the Markov Chain Monte-Carlo (MCMC) to converge to the target distribution by sampling each word and iteratively adjusting the model parameters based on the word topic. This is accomplished using Equation 2.18 where a variable is sampled, conditioned on the other variables. The first ratio in the sampling Equation 2.18 uses word frequency by total words, $\hat{\phi}_j^{(w)}$. The second ratio divides word-topics by total words in that document $\hat{\theta}_j^{(d)}$. Both leave out the current assignment of z_i and the results are used to randomly choose a new topic according to the current distribution. This process iterates until the log likelihood stabilizes [48].

$$P(z|w) = \frac{P(w|z)P(z)}{\sum_z P(w|z)P(z)} \tag{2.11}$$

Each term is expanded and assumed to have the indicated distribution, given as

$$p(w_i|z_i, \phi^{(z_i)}) \sim Discrete(\phi^{(z_i)}) \tag{2.12}$$

$$p(\phi) \sim Dirichlet(\beta) \tag{2.13}$$

$$p(z_i|\theta^{(d_i)}) \sim Discrete(\theta^{(d_i)}) \tag{2.14}$$

$$p(\theta) \sim Dirichlet(\alpha). \tag{2.15}$$

Based on the definitions of a Dirichlet and discrete distribution, the Bayes formula in Equation 2.11 is expanded into Equations 2.16 and 2.17.

$$P(w|z) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)} \tag{2.16}$$

$$P(z) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^M \prod_{d=1}^{M} \frac{\prod_w \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_\cdot^{(d)} + T\alpha)}. \tag{2.17}$$

These equations are simplified by integrating out ϕ and θ [48], thus computing the joint distribution $P(w, z)$. $P(w, z) = P(w|z)P(z)$ and ϕ and θ are isolated to each term, allowing us to integrate each out separately. Eliminating constants with respect to a change in the topic z results in Equation 2.18.

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\bullet)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha} \tag{2.18}$$

Solving for $P(z|w)$ must be approximated by using MCMC [48] to iteratively solve for z. Theoretically, this method could converge to the true model given an infinite number of samples. We choose a number of iterations that provides effective latent topic extraction without creating undue processing overhead.

2.4.5 Latent Dirichlet Co-Clustering.

While LDA has demonstrated wide-spread application for text clustering, it still assumes topical probability distribution across the entire document. Shafiei and Milios have developed a technique called Latent Dirichlet Co-Clustering (LDCC) [110] that models both document topics and word topics jointly. Where most probabilistic topic models capture the correlation between words, few are able to capture the correlation between topics. By doing so, one can decrease the word-topic cluster down to an applicable subset of a document (paragraphs or words, for instance), then model the overall topic of the document as a cluster of word-topics. This technique may facilitate image semantic tagging using local topics as clues, feeding into image

clustering.

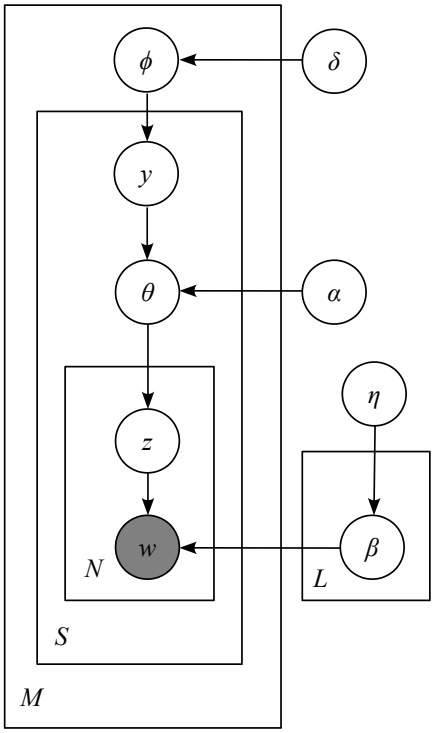


Figure 2.6: Latent Dirichlet Co-Clustering Model [110] Plate Diagram.

Figure 2.6 defines the plate diagram for the LDCC model, adding the S plate for word-segment topics, assuming that the document-topic mixture ϕ is a mixture of the word-topics θ [110].

2.5 Multimodal Fusion

Multimodal fusion research attempts to create models that represent multiple data domains concurrently, such as images or text. Very little multimodal research has taken place with latent topics [22]. A few areas have seen success in multimodal fusion, such as work that ties video to audio. Hauptmann and Christel [53] found that combining video with other data almost always improved retrieval performance, while multimodal fusion also successfully identified scene changes [121] and story segmentation [139] along with improvements using captions when available. Shafiei,

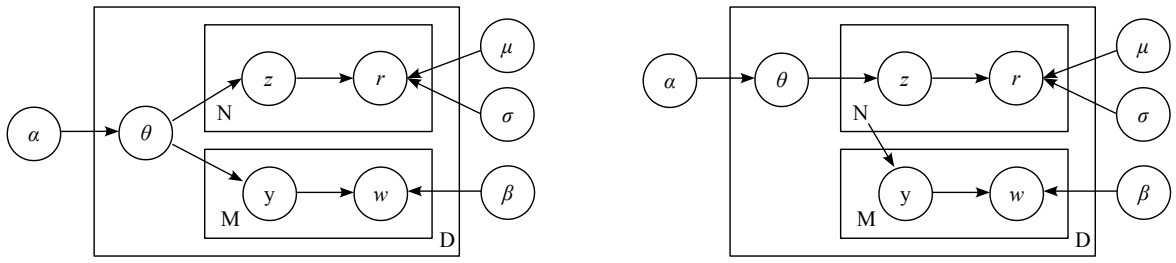
et al. further extended this research by topically segmenting a document corpus by paragraphs and sentences. Still others draw on a joint word-image modeling approach [78] [7] [60] [13] [70] that lays the foundation for the research in this dissertation.

This section sets the final foundation for our data mining method. The hypothesis is that both automated image annotation and latent topic extraction can be improved using multimodal fusion, similar to video retrieval. We first cover one of the dominant forms of image and text multimodal fusion: multimodal LDA. We then discuss some of the other methods for using image context to assist with annotation. Finally, we address a word balance problem that must be resolved prior to successfully using image tags to improve latent topic extraction.

2.5.1 Multimodal LDA.

Images often provide valuable context towards document topics and, similarly, text topics surrounding images help identify semantic labels for images. A system that can successfully exploit this synergy may improve both CBIR and latent topic extraction. Most data mining clustering research is focused on one particular domain, such as solely on text or images. There is, however, a growing body of research interested in drawing out latent variables from a corpus made up of both words and images. Initial research into multimodal representations [7] in data mining used soft clustering to arrange co-occuring text and image blobs into a hierarchical tree structure.

Blei and Jordan [13] tag regions of an image based on their mixture clustering. The regions can be assigned topics relative to the word topics. Their Generalized Mixture Latent Dirichlet Allocation (GM-LDA) model provides a multimodal approach to link words and image regions under the same generative latent topics. In this model, it is assumed images and captions are generated individually from the same document corpus (see Figure 2.7(a)). The Correspondence LDA (Figure 2.7(b)) model further



(a) GM-LDA.

(b) Corr-LDA.

Figure 2.7: LDA Image Variant Plate Diagrams.

expands this concept, assuming regions produce captions. Since then, other models have leveraged and modified this approach. This includes query logs, captions, and surrounding text linked to image blobs [78] cross-media relevance models [60], news articles [36], twitter topics [96] and life patterns using cell phone data [32].

Instead of using Gaussians to directly represent image blobs, Barnard, et al. [7], uses k-means to cluster the blobs into discrete labels. They then apply Dirichlet priors to both the word and image blob multinomial distribution. The words and image regions are linked using discrete-translation. Image regions still generate word annotations, but the multinomial distribution representing the blob labels provides greater granularity than Gaussian-based topic distributions. This model is useful for linking words to image blobs but grows in complexity as the number of words potentially associated with each image blob grows. Incorporating the paragraphs of text surrounding an image would greatly increase the complexity of this approach.

All these approaches assume semantically relevant information can be gained using the co-occurrence of words and specific image blobs or regions. Unfortunately, this may not provide adequate precision for an effective generalized classifier. The subject of most images typically lies in the center with a large surrounding region

of background. Most discussion about the image focuses on the subject rather than the background images, yet these regions would dominate most images. For instance, text surrounding an image of a boat would rarely mention sky or water but would typically talk about boats, naval strategy, cruise information, or similar topics. Images of beaches would also include significant portions of sky or water, yet would be surrounded by very different topics.

2.6 Word Balance

Limiting context to that surrounding an image, then pruning based on relevance, creates a new problem. Resulting tags are highly relevant since they relate directly to an image, yet their low frequency means they have little impact on overall topic extraction. One of the limitations with the Gibbs Sampling technique is the tendency to become trapped at the mean of a data distribution. This happens when the variance is small enough, especially when initial data points are chosen poorly. Often, it requires either restarting the chain or running it for extended periods [138]. Words with high frequency have a tendency to create distributions with high probability and small variance, not only preventing a topic from isolating semantically interesting words but drawing other words into it that may not belong due to the probabilistic strength of the topic. Additionally, some models are sensitive to probabilistic islands [138] due to high frequency, low standard deviation data clusters. During the Monte-Carlo simulation, if these high frequency spikes are encountered early, they can result in the model becoming "stuck".

These pitfalls with Gibbs LDA only complicate image-generated tags. If two images both produce 'water', yet one is about a recreational cruise in an advertisement and the other a swamp in an environmental document, we want these to cluster into separate topics. Additionally, accurate tags for an embedded image often carry

higher relevance than a single word embedded elsewhere in the document. Clearly, the low frequency of image-generated terms will have minimal impact on final topic distribution relative to their importance.

2.6.1 Word Balance Solutions.

Raw document text provides poor material for clustering and typically must be smoothed. Originally, LDA required not only cleaning up of some words, but changing zeros to very small numbers to do parameter estimation [14]. Commonly-occuring words, such as 'the', 'is', 'at', or 'on' will be represented by high frequencies within a document corpus. One solution is to simply remove these stop words prior to model formulation. Additionally, low frequency words do very little to help clustering or, even worse, actually prevent effective clustering. In many cases, low frequency words, sentences, or documents are removed prior to training the model. While common stop-word lists exist, defining these words can be rather ad-hoc, given there is at least $\binom{n}{x}$ stop-word lists where n is the number of words and x is the number of stop words [135]. Finding an appropriate subset can be challenging.

Another solution, defined by Jones, weighs terms based on collection frequency [65]. Coined Term Frequency - Inverse Document Frequency (TF-IDF), it penalizes words that appear in numerous documents, instead highlighting those that appear often in only a few documents.

$$tf - idf(t,d) = tf(t,d) * idf(t) \tag{2.19}$$

$$idf(t) = log\frac{|D|}{|\{d : t \in d\}|} \tag{2.20}$$

One could also eliminate stop words, then use TF-IDF to further refine the re-

maining words. Comparison testing demonstrated some success using this technique with Self-Organizing Map clustering. However, TF-IDF would preclude Gibbs Sampling from being effective since Gibbs Sampling relies on discrete frequency counts. There are, however, other techniques that allow for normalized frequency or TF-IDF in latent topic clustering. The Spherical Admixture Model (SAM) [97] is a Bayesian topic model that represents documents on a high-dimensional spherical manifold and can even assign negative weights to terms. While it did perform well with their test set, it is more complex and requires extra calculations to extract topics.

Recent research [124] indicates that fixing symmetrical priors for document-topic probability calculations prevents LDA from handling stop-words effectively. They have demonstrated that, by computing an asymmetric prior from the data, one can get stop-words to cluster into a few topics. This leaves the remaining topics full of semantically-significant grouping of words. Comparison testing with asymmetric priors using the Latent Dirichlet SEGmentation (LDSEG) [109] indicates significantly increased algorithmic complexity involved in recalculating priors.

Another method involves stemming the words to their root form. Originally defined in 1974 [25], a number of variants have been produced since that time [59]. While it does decrease the dimensionality of the documents, stemming has had some indication [51] that it performs poorly in retrieval tasks.

Frequently occuring words could be distributed in the form of n-grams. This is based on the idea that the words 'white house' may have vastly different contexts in a political versus real estate article. Merely storing the words 'white' and 'house' may miss valuable context that can be used to help cluster documents. The basic form of an n-gram consists of a bigram relationship between two adjacent words. By concatenating words, bigrams, trigrams or higher order n-grams can be produced. One user of the n-gram model [131] claims it produces reports that are more inter-

pretable than the standard unigram model. This has the added advantage that it decreases word frequency for those commonly-occurring words since they would be paired with another word that may lend higher significance to the pair (or triplet). It does, however, increase the complexity of the model and in comparison testing, produced worse results over unigram models.

Another potential solution [135] involves LDA term weighting. Instead of using word token frequency as in Equation 2.18, they use a weight for that word. While the actual term weight used could be based on any number of weights, they use point-wise mutual information between the term and the document, specified in Equation 2.21. This minimizes the impact of high frequency words that are spread out over a number of documents. While promising, this technique does not help to emphasize low frequency image-generated words over low frequency standard words.

$$m(x_i, d) = -log_2 \frac{p(x_i|d)}{p(x_i)} \tag{2.21}$$

Blei, et al. modified LDA slightly [12] to incorporate a supervised response variable into the generative model. For instance, a list of movie reviews could develop topics based on movie genre or whether the review indicated the movie was good or bad. While not directly applicable, a semi-supervised approach could be applicable, emphasizing the image text results to determine the latent topics within the document, skewed towards image-word / document-word matches.

2.7 Iterative Automated Image Annotation Improvements

A number of papers demonstrate improvements can be made by utilizing outside information. Jeon, et al. [60] used the Cross-Media Relevance Model (CMRM) to improve image annotation by linking image blobs with words. This was expanded using Random Walks [125] and a Markov chain model [126] coupled with the CMRM

to iteratively improve image annotations using captions. Experiments using CMRM tend to rely on a set of manually annotated images to train a system in identifying unknown images, thus would have difficulty extracting annotations where some words may be irrelevant to the image itself. Additionally, the trained model does not take into account local context when assigning annotations. Likewise, the nearest neighbor approach [82] also trains on words without local context. Feng and Lapata [36] used the continuous relevance image annotation model, coupled with LDA trained on words in news articles to improve automated image annotation.

2.8 Evaluation

A number of methods exist for evaluating the efficacy of text-based clustering. A common technique is to use precision and recall to measure the ability of a system to return relevant documents to a query. Recall measures the number of relevant documents retrieved for a query out of all documents in the corpus (Equation 2.23), while precision measures the number of relevant documents returned out of those retrieved (Equation 2.22). Since precision and recall by themselves are prone to return good results for bad performance (recall is 100% if all documents are retrieved for a query), the F-measure uses a ratio of recall and precision (Equation 2.24). Assuming one cluster per class, F-measure provides an effective estimate of performance. However, if the number of clusters and classes is significantly different, other measures are needed [137].

$$precision = \frac{|Relevant \cap Retrieved|}{|Retrieved|} \tag{2.22}$$

$$recall = \frac{|Relevant \cap Retrieved|}{|Relevant|} \tag{2.23}$$

$$F = 2 * \frac{Precision * Recall}{\beta^2 * precision + recall} \tag{2.24}$$

While Normalized Mutual Information (NMI) may provide a good measure of the model's clustering, F-measure demonstrates how well each model generalizes to new data. Since the overarching goal of this research is a generalized, robust classifier, F-measure is primarily used while precision and recall are given to provide additional information.

2.9 Large Corpus Data Mining

Part II investigates the efficacy of latent topic models on real world data in a digital forensics context. Therefore, a discussion about topical literature would not be complete without discussing the field of research in large corpus search or digital forensics. Traditional digital investigations using keyword or regular expression searches is approaching the cognitive limitations of a human analyst [9]. Proper keyword selection can be instrumental in the success of an investigation, requiring experience and potentially a bit of luck. Since two humans will select the same word to describe an object less than 20% of the time [39], a target document may have a relatively low chance of being selected. With data generated by multiple users, such as a network server or disks retrieved from a company, keyword searches must either include a variety of potential keywords and increase the false positive rate or constrain their keywords, decreasing the recall rate.

One solution to the keyword problem is to create a keyword list in an intelligent manner that, on average, will produce better results. Du, et al. [30] attempted this using WordNet and Latent Semantic Indexing to prune high-noise keywords. Probabilistic Latent Semantic Analysis (pLSA) along with several other clustering techniques can be used to identify semantically or topically similar documents [5].

Many of these techniques have been used in forensics analysis, such as the work by Beebe et al. [10] that successfully clustered query results using self-organizing maps. Most of the techniques used in this field, however, test against small, artificial, or single disk test data that only partially represents a real-world disk. Additionally, techniques for analyzing data are designed and tested against either small data sets or large, and fail when presented with the opposite data [41].

2.10 Test Data

Testing algorithms that use both text and images require a data set containing documents with sufficient embedded images. Truth information is useful since it provides the ability to measure algorithmic performance using precision and recall. The U.S. Patent and Trademark Office (USPTO) offers a large corpus of patents grouped by category [2]. Each file is in XML format and many include embedded images. Initial testing with the USPTO database found the images difficult for the automated image annotation algorithm to interpret. This was largely because most images were diagrams or close-ups of a mechanical part. The GovDocs corpus [40] provides over one million files from various government entities. Many have embedded images, though like the patent documents, they tend to be engineering diagrams, flow charts, or close-ups of machines. In addition, it lacks category information.

The INitiative for the Evaluation of XML Retrieval (INEX) 2007 Wikipedia data set [28] offers a large corpus of Wikipedia documents manually labeled with overlapping categories. The English version includes over 600,000 documents on a wide variety of topics. Documents tend to be focused and images tend to be highly relevant to the document topic. This makes it ideal for testing algorithms using a clean set of data, but does not represent the noisy data a forensics analyst might encounter.

The forensics analyst usually deals with large amounts of unstructured data. This

may include a single hard drive, or may contain hard drives, cameras, smart phones, or other storage devices. Testing large, cross-drive analysis against real-world digital forensics data is rarely possible since mixing evidence between cases is problematic [41]. To help solve this problem, Garfinkel developed a corpus of hard drives purchased off the open market with a variety of real user data. He used them to test a Cross-Drive Analysis technique [41] using lexigraphic data to identify real-world drives heavy with financial information, e-mails, or other information with defined formats. It provides a highly realistic set of data to assess algorithms using multiple data domains.

2.11 Summary

This chapter presented the background research pertinent to the methods used in this dissertation. It first discussed automated image annotation techniques. Related techniques, such as ensemble directors and image segmentation techniques. It then discussed text mining and comparison techniques that provide the foundation for the SWLDA model discussed in Chapter V. This includes discussions on multimodal latent topic extraction, along with the complications of some of those methods. Finally, large corpus search is reviewed along with potential solutions for improving document search and retrieval accuracy.

III. Part I: System Overview

Data mining and digital forensics for multimedia data is challenging, partly due to a lack of structure within the data content and partly a result of limited research within multimodal data mining. Fusing data mining techniques may produce better accuracy for user queries than individual methods. While a comprehensive data mining fusion tool would bring together audio, video, still images and text, this research limits its scope to still images and text. By fusing methods together between these two digital domains, improved accuracy may result in addition to a more robust search capability.

This chapter provides the details of the automated image annotation and latent topic iterative model. It discusses the overall model, assumptions, and justifications. The three stages of the model are introduced and are further detailed in Chapters IV through VI along with their testing and results. Chapter VII assesses what aspects of this model, if any, can be run on a set of real-user hard drives.

Figure 3.8 diagrams the full model process. The inputs to the system include a document corpus with embedded images. The document parser extracts text and image information, removing stop-words and passing the individual words to the Stage I automated image annotation algorithm. In Stage I, the model attempts to leverage the text surrounding an embedded image as context to improve basic image annotation. The proposed model is demonstrated to provide improved annotation when compared against the Automated Linguistic Indexing of Pictures in Real Time (ALIPR) image annotation system [73]. In Stage II, those image tags are used as super-words in the SWLDA model to improve latent topic extraction from the document corpus. The posterior probabilities within the SWLDA model facilitate document retrieval and topical browsing. In Stage III, these posterior probabilities help prune topically-insignificant words from the image annotation and the surviving are words used to

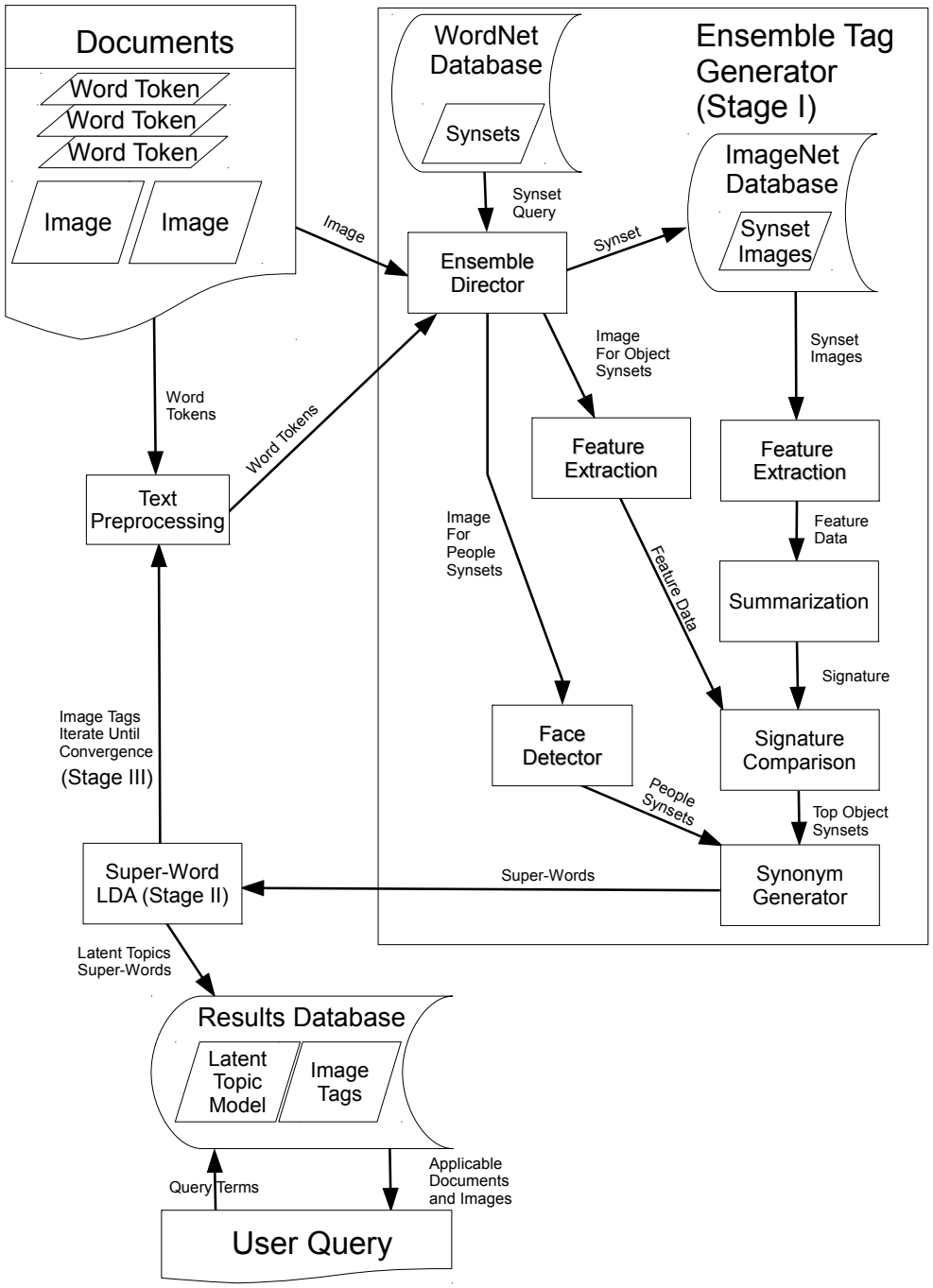


Figure 3.8: Automated Image Annotation and Latent Topic Iterative Model.

expand the breadth using WordNet hypernyms and hyponyms. This can be fed back into Stage II to refine latent topic extraction and image annotation further. Part II

of this research evaluates the effectiveness of generative topic models on a real world non-curated data set.

3.1 Document Processing

The documents within the Wikipedia XML corpus are read using the libtinyxml library in Java on a Ubuntu linux machine. The raw text is parsed and loaded into a MySQL database and the text location recorded. Images are placed in a directory on the disk with a database entry providing the location of the file. All images are loaded using the OpenCV library. All subsequent processing of images and text are conducted on the database.

IV. Stage I: Automated Image Annotation

The context-driven image processing approach draws on the advantages of both generalized and specialized classifiers in a novel way using contextual information surrounding an image. It automatically assigns image annotations to embedded images using local context, WordNet, and signatures generated from ImageNet. This chapter describes stage I of the Automated Image Annotation and Latent Topic model. It first discusses the methodology and experimental design, then summarizes and analyzes the results.

4.1 Methodology and Experimental Design

Stage I leverages WordNet synset information from local text context to prompt an ensemble director that selects the best classifier based on intrinsic knowledge of that synset. ImageNet images, organized by synset, then provide a new capability to build signatures representative of that synset and compare them against a query image. ImageNet includes over 14 million images embedded into a hierarchical structure based on WordNet with over 21,000 synsets that can be used for classification. Many of these images include bounding boxes and human annotated attributes that provide a valuable resource for image tagging research. While Leong and Mihalcea [72] used image features from ImageNet to measure word relatedness between synsets based on their image similarity, we are not aware of any similar use of ImageNet to compare against unknown images within an ensemble director.

The method, shown in Figure 4.1, can be applied to any document corpus with images surrounded by text. The document structure is parsed and the image caption, paragraph before, and paragraph after the image extracted, as indicated by the Documents block in the upper left of the figure. Stop-words are removed by the "Text

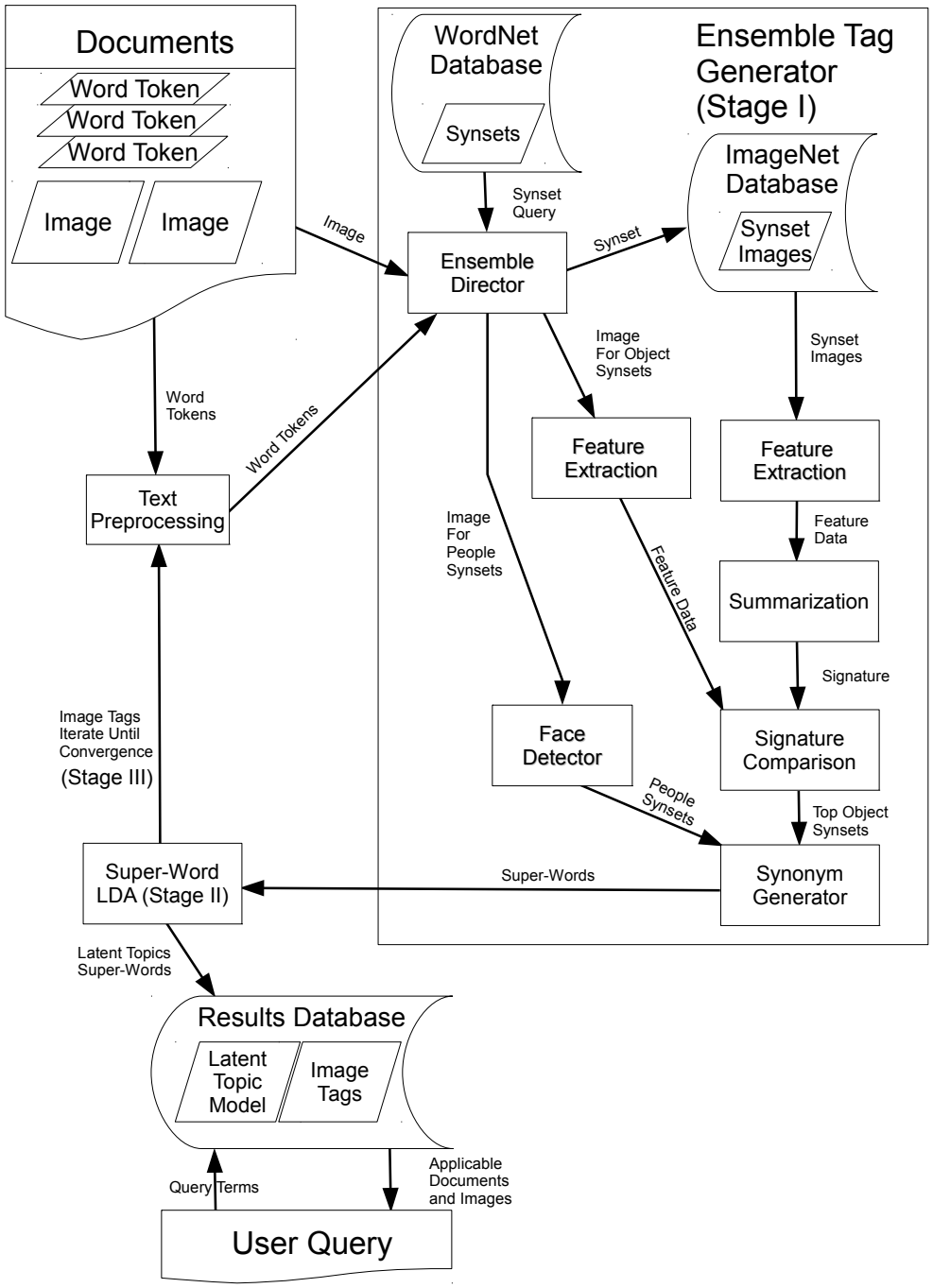


Figure 4.1: Automated Image Annotation and Latent Topic Iterative Model.

Pre-processing" step and a list of noun synsets are generated for each remaining word. ImageNet has a large number of synsets in its hierarchy; however, not all have images

associated with them so not all synsets can be tested. Additionally, only nouns have associated images even though it may be relevant in some cases to tag images with adjectives or verbs. For this reason, not every word in the surrounding paragraphs can be rank ordered.

The "Ensemble Director" first analyzes the image features using a histogram to detect if the image is a graph or line art. These images are skipped, though future research could incorporate Optical Character Recognition (OCR) to extract information. The synset of the image is then extracted and, if its root synset is person related, then the image is sent to the face detector module (see Section 4.3.3). If a face is detected, that word is marked as potentially applicable. If the synset falls under any other root synset, the image is segmented using Efficient Graph-Based Image Segmentation [35] in the "Feature Extraction" step and the mean vectors calculated. A normalized histogram is generated from these mean vectors, both for the unknown image and a subset of images from the ImageNet synset. The histograms are compared using Earth Mover's Distance (EMD) (see Section 4.3.4) in the "Signature Comparison" step and the words rank-ordered from least EMD to greatest. The resulting words can then be used to tag the images based on a calculated EMD threshold or fixed number of tags. There currently is no way to distinguish between words related to people–either faces are detected and all people words are accepted, or no faces are detected and the people words are discarded.

4.2 Text Preprocessing

Raw text includes a number of frequently occuring words, such as 'a', 'and' and 'it' that do not provide significant information to the algorithms in this dissertation. This section discusses the problems considered when designing the methods in this chapter, as well as the specific text pre-processing steps necessary to produce good

results.

Image captions and surrounding text can be used to help refine image tagging. Unfortunately, words by themselves are still ambiguous without their surrounding context. For example, use of the word 'plane', meaning an aircraft, or 'plane', meaning a tool for smoothing wood can be identified by how it fits into the paragraph. This work hypothesizes that significant differences exist in the image features of aircraft and tool 'planes' that permit rank-ordering tags for an unknown image without relying on paragraph context. Additionally, the text surrounding an image will help constrain classifier category overlap and allow highly-relevant words to distinguish themselves. Nouns tend to be more descriptive of images while adjectives, verbs, and even adverbs may contribute to effective image annotation. For example, 'white' and 'house' mean something different than 'white house'. While digrams and trigrams provide opportunity to expand this research and may improve results, we constrain complexity by limiting the research to single word instances.

By combining text, object recognition methods, and statistical pixel analysis, context clues can be leveraged to develop a generalized image annotator. The challenge is in limiting the search space and categorical overlap. The text surrounding an image may provide contextual prompting, allowing for careful selection of appropriately-specialized classifiers. Unfortunately, specific concepts are often represented by a number of words and each word may have several meanings or synsets. Accurately mapping to the intended meaning requires complex lexical analysis or a search across each possible meaning. WordNet [34] is a lexical database that groups words into distinct concepts called synsets. Additionally, it includes a hierarchical hypernym/hyponym relationship between more general and specific forms of a word. This allows the words surrounding an image to provide a robust contextual prompting based on meaning instead of text. To select the most applicable meaning based on the sur-

rounding words, Popescu, et al. [95] utilize the WordNet hierarchy to group images of placental animals within their hypernym/hyponym synsets as chosen by the user. They compared pixel statistics of known images within the synset and unknowns to accurately select from sets of unlabeled images. As expected, the more specific terms (e.g. Brown Swiss cow) perform better at finding similar images than more generalized terms (e.g. Bovine). Google research [79] [64] has also leveraged contextual information via hyperlinks and user selection along with image features to aid image annotation.

Word distance from the image is an additional concern. This research hypothesizes that the closer a word is to the image, the more likely it is to be pertinent. For the proof of concept, we simply use the image caption and the paragraph before and after the image. However, results may be improved by a smarter extraction algorithm. A distance threshold may account for the disparity in paragraph size while intelligent parsing could look for references, e.g. 'Figure 1', and extract the text surrounding that reference. The goal, however, is to limit word extraction to maximize the chance of extracting pertinent words rather than all words, which may be less relevant.

Words are also limited to nouns with available ImageNet synsets. Some synsets only have a handful of images associated with them and are also ignored. During initial experimentation, synsets with fewer than ten images performed poorly so that number was chosen as the minimum threshold. Stemming was deemed not necessary since WordNet already recognizes alternate forms of certain words.

4.3 Ensemble Director

An ensemble director is used to gain higher accuracy by using the synset hierarchy to drive the decision tree, as outlined in [91]. The ensemble director detects graphs or clip-art, people, and all other images, handling each differently. While ImageNet

does contain a wide variety of images associated with people, it is unlikely an image classifier will be able to distinguish between an image of a medical doctor and one of a nurse. The ensemble classifier overcomes this limitation, and is outlined in Algorithm 3.

Detecting graphs, faces, and similarity with other images is challenging given that images consist of pixel values without obvious structure. The techniques used by the ensemble classifier require the pixel information be summarized and attributes, such as frequency, must be extracted. This section first discusses the methods used to select image features for the ensemble director. It then discusses the modules used to identify graphs, faces, and image similarities.

Algorithm 3 Ensemble Director Algorithm

```
Input: Image, synset
Output: Rank ordered super-words
if Image is a graph then
  Discard Image and Exit
end if
for all Synset of words surrounding the image do
  if synset is people-related then
    if Faces are detected in the image then
      Add the synset as a super-word
    else
      Discard Super-Word
    end if
  else
    Compare image against synset image signatures and calculate Earth Mover's
    Distance
  end if
end for
Convert image color space to a CIE LAB histogram
Identify the three largest histogram spikes
Calculate the standard deviation around the spikes
Return top three synsets ranked by Earth Mover's Distance or People-based words
```

(a) Original Image (b) LAB Color Space. (c) LUV Color Space. (d) HSV Color Space.

Figure 4.2: Color Space Comparison Using k-means Pixel Clustering.

4.3.1 Image Comparison Feature Selection.

This section describes the development of the feature extraction, summarization, and signature comparison techniques that are used in the face detection, graph detection, and ImageNet-based signature comparison modules. Features were selected using a mixture of comparison testing and intuitive analysis. A number of methods were tested for their ability to cluster images and extract accurate similarity measures. These include color, frequency, shape, size, and location. The results of the feature selection testing is described in this section.

4.3.1.1 Color Space.

As discussed in Section 2.1.1.1, each color space has its benefits. Three of the most prominent include the CIE LUV (good for glowing colors), CIE LAB (good for reflexive colors), and the HSV (accurately represents painting colors). The images from left to right are: the original, the CIE LAB color space, CIE LUV, and finally HSV.

Figure 4.2(a) through 4.2(d) illustrate the differences between the color spaces. A

k-means average was taken for the pixel values in each color space, then the pixels changed to the mean of each cluster. The CIE-LAB color space isolates the horse best, while the CIE-LUV color space image breaks apart the horse into dark and light regions. The HSV color space breaks the horse into multiple blobs that do a poor job of representing the actual horse. Based on widespread use in the literature [19] [74] [36] [62] and success in comparison testing, the CIE-LAB color space is used for this research.

Each image is loaded from its file format into the RGB color space where it is then converted to the applicable space. The color vector (e.g. $C = [L, A, B]$) can then be summarized using either k-means or expectation-maximization, then used in the image clustering algorithm.

4.3.1.2 Frequency.

Frequency, as illustrated by the Daubechies wavelet decomposition of the bus in Figure 4.3, has demonstrated improved automated image annotation during comparison testing. The wavelet is computed over the CIE-LAB color space. Previous experimentation demonstrated that the L dimension produced far better results than either the A or B spectral dimensions. Adding either of these wavelets to those produced by the L dimension tends to generate less precise clustering than just the L dimension alone.

As with color, frequency can be summarized into representative means and used to cluster images. Reducing each image into two filter iterations with all three frequency bands produces better clustering than color alone. Utilizing only the first iteration of the filter tends to produce a poor clustering signature, while using a signature made of the first two iterations demonstrates a much higher fitness. Averaging the horizontal, vertical, and diagonal frequencies together for each iteration improved

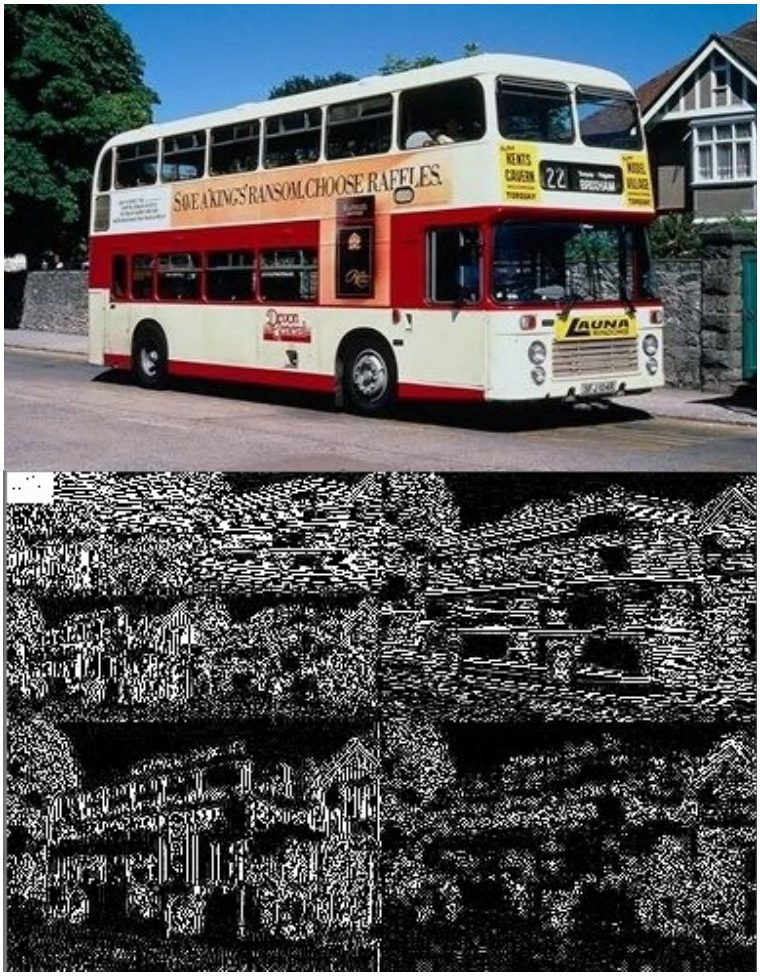


Figure 4.3: Frequency Wavelets - Bus Image.

clustering further during comparison testing.

The feature vector used in the image processing algorithms consisted of the three dimensions of the LAB color space. It also included the average frequency between the horizontal, vertical and diagonal wavelets for the first and second application of the wavelet filter.

4.3.2 Graph Detection.

While future versions of this ensemble director may handle clip-art and graphs, the current version does not. It uses a technique adapted from [95] to detect graphs using

standard deviation around the largest lightness histogram peak. That method, however, resulted in a number of false positives where images had significant amounts of white or black in them. Hence, it was modified to include the average standard deviation of the top three histogram peaks. This prevented a single color from dominating and greatly improved graph detection. Graphs are then discarded as untaggable. The pseudo-code is defined in Algorithm 4.

Algorithm 4 Graph Detection Algorithm

Input: Image, max_ratio
Output: boolean value indicating whether it is a graph
Convert CIE-LAB image pixels into a color space histogram using 100 buckets per channel
Identify the three largest histogram spikes (s_1, s_2 and s_3)
Calculate the standard deviation around the spikes up to three buckets away
for all spike pairs s_i and s_j **do**
  if $s_i/s_j < max\_ratio$ **then**
    Eliminate s_i
  end if
end for
if (The spike magnitude / largest spike magnitude) > threshold of 0.2 **then**
  Add the spike standard deviation to the average standard deviation
end if
if The average standard deviation is greater than 0.3 **then**
  Return true.
else
  Return false
end if

4.3.3 People-Based Synsets.

Determining if people exist within an image is complex, with a number of different methods available [66] [122] [123] [108] [105]. Comparison testing demonstrated that looking for skin colors is found to produce excessive false positives. Face detection, on the other hand, provides more accurate results. People detection utilizes a boosted cascade of Haar-like features [123] from the OpenCV 2.4.6 pre-trained alternate frontal

face detector. This accurately detects faces when they are dominantly displayed in the image, though has reduced accuracy with side profiles and small images. This is likely acceptable since, when people do not dominate an image, they are less likely to be the focus of an image.

4.3.4 Remaining Synsets.

ImageNet provides an assortment of images for select synsets that consist of images from all sub-categories. For instance, the synset for 'dog' contains a wide assortment of dogs while 'Great Dane' contains only that particular breed. This helps when generating a signature since the signature for 'dog' will be broader than the narrower 'Great Dane' subset.

Before a signature is generated, the features are extracted from the image and summarized similar to the method outlined in Section 4.3.1. Once generated, the features are reduced to representative samples to better highlight the defining traits of an image set. This is accomplished through mean vectoring of the features within a region. The Efficient Graph-Based Segmentation [35] produced image regions whose vector means represent features common to all images of a particular synset. Efficient Graph-Based Segmentation utilizes a ratio of region size to threshold node merging based on edge gradient. Unfortunately, this requires parameter tuning in relationship to image sizes. In an image with numerous pixels, the threshold shrinks as pixel size increases, resulting in a high number of regions. Because of this, all images are resized to a width of 400 pixels while maintaining the original aspect ratio.

By calculating the region mean vectors for each image, a histogram of images are assigned to a particular synset. Ideally, this histogram includes a representation of those synset elements that occur frequently versus another synset. Comparing the query image to the synset signature must account for different primary subject

Table 4.1: Image Dimensions

L	Lightness component of CIE-LAB color space
a	'a' frequency of CIE-LAB color space
b	'b' frequency of CIE-LAB color space
DB1_AVG	Daubechies Horizontal/Vertical/Diagonal avg high frequency
DB2_AVG	Daubechies Horizontal/Vertical/Diagonal avg medium frequency

backgrounds or sizes. Any similarity measure must detect general patterns while not being confused by the noise within the domain. The EMD provides a robust comparison algorithm that has performed well in related image retrieval testing [101]. It represents the amount and distance of 'earth' or histogram area that must be moved to convert one histogram to the other. This provides a numerical quantification of differences between an image's histogram and a synset's representative histogram.

The implementation of the EMD includes several preprocessing steps. First, to minimize noise impact within the synset histogram, all histogram buckets with fewer mean vectors than the threshold are dropped. This highlights frequently occurring elements of that synset while minimizing background regions that do not appear consistently throughout the images. Second, comparison testing determined that using too many images to represent a synset averaged out key traits. Additionally, too few histogram buckets lost data while too many created a sparse signature. An image count of between ten and forty appears to produce the most accurate results, while synsets with less than ten images are ignored. A histogram bucket size of 26 best balances detail with sparsity.

4.3.5 Image Processing Modules.

The final modules use Efficient Graph-Based Segmentation with Mean Thresholding to summarize the image features into mean vectors. Images are normalized to a width of 400 pixels while maintaining aspect ratios. The vectors were used to populate a histogram with 26 buckets per dimension using the three LAB color space

Table 4.2: Wikipedia Test Categories.

Category	Documents	Images	Pruned Images
Skyscrapers	258	83	75
Aircraft	1877	501	441
Vegetables	324	62	56
Armored Fighting Vehicles	143	48	47
Dogs	813	231	215
WWII Ships	788	26	22
Flowers	164	37	34
Elephants	258	19	18
Sailboats	132	24	23
Mountains	2252	305	260

channels and the average Daubechies wavelet between horizontal, vertical and diagonal wavelets. The first and second iteration averages were each used. The dimensions are given in Table 4.1.

4.4 Evaluation

Testing of the algorithm used the 2007 INitiative for the Evaluation of XML Retrieval (INEX) dataset [28]. The INEX dataset includes more than 600,000 documents from Wikipedia, many with embedded captioned images. Each document is labeled with overlapping topic categories, providing a robust mechanism for selecting a range of document topics. In addition, Wikipedia articles tend to be topically related to the images embedded within, helping to highlight poor results as algorithmic problems rather than inconsistent data. Other real world data repositories may include far noisier data and not perform as well. Some of the images were too small while some looked like graphs. The number that were actually annotated is listed under the Pruned Images column.

The results are compared with Automated Linguistic Indexing of Pictures in Real Time (ALIPR) [73] due to its availability and widespread acceptance in the research community. Each image tested using this dissertation's algorithm was also sent to ALIPR and tagged. The top fifteen words produced by ALIPR were then sent to a workforce of human analysts on the Amazon Mechanical Turk (AMT) website for scoring. Unfortunately, ALIPR does not provide an indication of the synset for each word, so definitions were not provided to workers.

The human analysts on AMT graded the image annotations on a scale of one to five, where 'five' indicates the word describes the image completely and 'one' indicates no correlation between the word and image. Each word annotation included the definition from the WordNet synset to remove ambiguity on the intended word form or synset. Five AMT workers were used to score each word, with the average score considered as the standard and disagreement represented by a standard deviation value.

This test compared two things. First, the automated word ranking should match as closely as possible to human rankings. In this research, this is limited only to those words both algorithms extracted from the document and were available in ImageNet. Second, the words should be specific enough to be useful descriptions. Describing a boat as a 'thing' may be highly applicable but too general to be useful.

Performance is measured as the average Amazon Mechanical Turk score from the top five words. In addition, specificity is used to determine how specific the top five words are. Up until recently, lexical analysis research has used the Term Frequency - Inverse Document Frequency (TF-IDF) to help determine specificity. Recent research [102] uses a mixture of WordNet attributes to define the 'importance' of a word, given its generality or specificity. They utilize a function of term weighting based on the number of senses in a word, the number of synonyms, the hypernym level, and the

number of children (hyponyms/troponyms). Since specific synset information is not available for ALIPR, this research utilizes an average calculation to determine best and worst-case specificity levels. To simplify this calculation, only the hypernym level is considered.

4.5 Results and Discussion

The Stage I results were calculated for each image and are listed in Table 4.3. The data provides the overall improvement of the context-driven algorithm over ALIPR. Since ALIPR did not provide the synonymous set for a particular word, three values must be calculated to estimate the actual value. First, the minimum specificity takes the least specific synset for each word and produces the worst-case scenario. The second value indicates the average of all possible synset specificity values. Finally, the third calculates the greatest possible specificity, or best possible case. As our algorithm provides synset information, this calculation was unnecessary for the context-driven approach, so those fields in Table 4.3 are marked as N/A.

Table 4.3: Method Test Results Comparison.

Method	Relevancy	Min Spec.	Avg Spec.	Max Spec.
Context-Driven	1.98	N/A	9.09	N/A
ALIPR	1.62	5.03	6.43	7.85

The data in Table 4.4 divides this performance into the various document categories. The third column provides the average relevance measure for each category while the fourth column indicates specificity. Some categories performed better than others. For example, documents about elephants had images that poorly represented the topic, such as images in the dark, covered in tapestries or lights, or consisting of charcoal drawings. The human confusion is evident as this category had the second highest standard deviations among AMT worker scores. Vegetables had the

largest standard deviation, demonstrating the difficulty that humans had in determining what kind of label a leafy plant or variety of squash receives. While the elephants were hard for the algorithm to annotate, it had fewer problems labeling vegetables, likely due to the consistent color and texture among common varieties. The skyscraper category scored low in specificity since it tended to use very generalized words such as tower, structure, and building. Images of 'WWII Ships' tended to be grainy black and white, taken from a distance or with high background noise.

Table 4.4: Relevancy and Specificity Scores by Category.

INEX Category	Image Count	Avg Relev.	Avg Spec.
Elephants	18	1.75	9.15
Mountains	260	1.76	8.44
Aircraft	441	2.07	9.59
Dogs	215	2.00	10.10
Skyscrapers	75	2.00	8.75
Sailboats	23	2.07	8.67
Armored Vehicles	47	2.00	9.61
WWII Ships	22	1.82	8.93
Vegetables	56	2.16	9.02
Flowers	34	2.11	9.25

Table 4.5 provides an example of the context-driven algorithm performance on the image of a B-1B bomber aircraft. Words drawn from the area surrounding the image help populate the table and the algorithm rank orders them based on their image features. The first column provides the ordered list of words generated by this context-driven algorithm and the second column has a very brief definition. The definitions had to be abbreviated due to space limitations, but they illustrate the difference between two variants of a word. The ordered top words generated by the ALIPR algorithm are in the third column. As expected, bomber scored near the top while completely unrelated words, such as 'ocean' and 'range' were pushed to the bottom. The word 'cockpit' appeared high on the list, though no cockpit was directly

Table 4.5: Sample Image Annotation Results - Aircraft Image.

Word (in order)	Definition	ALIPR words
cockpit	Area pilot sits	man-made
bomber	Aircraft that drops bombs	indoor
panel	Electrical device	photo
control	Mechanism	old
throttle	Controls throttle valve	people
throttle	Regulates fuel to engine	decoration
engine	Motor	decoy
range	Series of mountains	snow
ocean	Large body of water	ice
limited	Public transportation	winter

visible, likely due to color and texture similarities between the aircraft image and an actual cockpit.

Table 4.6 provides results from one of the more challenging images within the 'dog' category. The algorithm appears to match color and texture features common to dogs, humans, and a thoroughfare (street) from the ImageNet database and rank-ordered 'dog' and 'human' as one of the highest. The alternative definition for the word 'dog', meaning metal supports for logs in a fireplace, was ranked low due to different image features.

4.6 Summary

This chapter provided evidence supporting the hypothesis that local context can improve automated image annotation using simple image comparison techniques. An ensemble director uses the text surrounding an image to select an appropriate image

Table 4.6: Sample Image Annotation Results - Dog Image.

Word (in order)	Definition	ALIPR words
dog	Domestic dog	people
human	family Hominidae	man-made
street	Thoroughfare	sport
sign	A public display	car
street	Thoroughfare (variant 2)	cloth
control	Operates a machine	plane
sign	Advertising board	guard
retriever	Dog variant	parade
people	Group of humans	sky
blind	A protective covering	race
dog	Supports for fireplace logs	motorcycle

comparison algorithm. The product of the ensemble director is a rank-ordered list of possible word annotations. These annotations were demonstrated to have higher average relevancy and specificity than ALIPR, a popular generalized image annotation algorithm. The results from Stage I are used to improve automated topic extraction in Stage II, described in the next chapter.

V. Stage II: Super-Word Latent Dirichlet Allocation

Stage II tests the hypothesis that high-relevance super-words produced by automated image annotation will improve latent topic model parameter estimation. This, in turn, results in better topic prediction for unknown documents when compared to existing latent topic models. To test this hypothesis, image annotations produced by Stage I are used as super-words to estimate hidden parameters of a new latent topic model called the Super-Word Latent Dirichlet Allocation (SWLDA) algorithm. This chapter first discusses the model design and methodology, then describes the experiments. Finally, it describes and analyzes the results.

5.1 Methodology and Experimental Design

Latent Dirichlet Allocation (LDA) is a generative model that assumes documents and topics are generated from a model with hidden parameters (see Appendix B for an intuitive description). These documents and topics then generate the words of a document. If the model parameters are known, it is possible to draw inferences about unknown data. For instance, if one has a new document containing words and hidden topics that was assumed generated from the model, then the probabilities for the document topic can be assumed. Likewise, if the topic is known, a list can be generated of the most likely words for that document.

Unfortunately, the model parameters and topics are rarely known a-priori. Documents and words exist and can be used to estimate model parameters using various approximation techniques [14] [48]. Many documents contain both words and embedded images, yet LDA only estimates model parameters for the text, ignoring images that may be highly relevant. Some algorithms have attempted to draw in image blobs [13] [78], but only test the relevancy of words linked to blobs, not whether

blobs improve model estimation. Rather than using blobs in their raw image form, the SWLDA model proposed in this chapter first estimates word annotations for images, then uses those annotations as super-words to improve the model's ability to draw inferences on unknown data. Image annotations from Stage I are used to populate a set of super-words S over another set of words W without high algorithmic complexity or excess destruction of data.

The SWLDA model assumes two subset vocabularies W and S drawn from a common vocabulary V, where $W \cap V = W$ and $S \cap V = S$, but $W \neq S$. For the purposes of consistent notation, S typically would have a smaller size than W, but would contain words of higher relevance to the topic model. Each super-word is generated by documents and latent topics, similar to standard LDA.

Each word in the set S has greater topical significance than the set W but may not have as high a frequency. Due to their higher topical significance, we call the words in set S super-words over the regular words in set W. Like in standard LDA, each super-word is generated by documents and latent topics. While this model is designed and tested with a document corpus in mind, it should work for any generative entity with elements of stronger co-occurrence.

The plate diagram in Figure 5.1 is derived from the Bayes formula given in Equation 5.1. The SWLDA model is assumed to generate words, latent topics, and documents, similar to standard LDA. In this case, documents and topics also generate super-words with the same 'bag of words' assumption. Variables w and s represent the observed N words and S super-words generated by the model. These words and super-words are generated according to the multinomial probabilities ϕ and τ respectively. The multinomial parameters are themselves generated by Dirichlet distributions with priors β for words and η for super-words. Document topic probabilities z are generated by the multinomial distribution θ and those priors generated via the Dirichlet

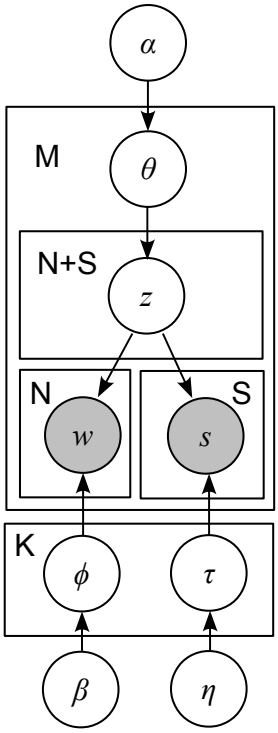


Figure 5.1: Super-Word Latent Dirichlet Allocation Plate Diagram.

distribution with priors α. Word w and super-word s generation are influenced by the document-topic probabilities.

$$p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) = \prod_{k=1}^{K} p(\phi_k; \beta) p(\tau_k; \eta) \prod_{m=1}^{M} p(\theta_m; \alpha) \left(\prod_{n=1}^{N_m} p(z_{m,n}|\theta_m) p(w_n|\phi_k) \times \prod_{i=1}^{S_m} p(z_{m,N_m+i}|\theta_m) p(s_i|\tau_k) \right) \tag{5.1}$$

Since only the words and super-words are known, the multinomial parameters θ, ϕ, and τ must be estimated using the known data. The goal is to solve for the unknowns using Equation 5.1. This cannot be accomplished directly but must rely on approximation techniques. Gibbs sampling offers a solution using a Markov Chain Monte-Carlo (MCMC) to iteratively approach the true model parameters. This re-

quires a sampling formula to determine word and super-word topic assignment similar to the standard LDA Gibbs sampling formula defined in 2.18. Starting with Equation 5.1, we aim to isolate a proportional equivalent with respect to changes in topic, or k.

$$p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) = \int\int\int p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) d\theta d\phi d\tau \tag{5.2}$$

The integral of the probability for the support of Equation 5.1 is given by 5.2. Conditional probabilities are considered independent of all non-descendants given their parents, so the integrals can be isolated with respect to θ, ϕ and τ. Normalizing constants are assumed in the resulting Gibbs sampling equation but not shown. Constants are removed and proportionality maintained with respect to a change in the topic k, resulting in Equation 5.3 (proof demonstrated in Appendix A). The Gibbs sampling proportional probability is based on the selected topic k, document d, and word e. The term f is a boolean term indicating either a word (γ_w) or a super-word (γ_s). The term $c_{k,d,e,f}$ represents the word count of the word indexed by k, d, e, and f. Likewise, $z_{d,e,f}$ represents the topic assignment at that particular word ($1...K$). Probabilities must be normalized.

$$p(z_{d,e}|z_{-(d,e)}, w, s, \alpha, \beta, \eta) \propto \begin{cases} t_{d,e} \in W: & \dfrac{(c^{-(d,e,\gamma_w)}_{z_{d,e},d,\bullet,\bullet} + \alpha_{z_{d,e}}) \times (c^{-(d,e,\gamma_w)}_{z_{d,e},\bullet,w_{d,e},\gamma_w} + \beta_{w_{d,e}})}{(c^{-(d,e,\gamma_w)}_{z_{d,e},\bullet,\bullet,\gamma_w} + V\beta)} \\ & \times \dfrac{(c_{z_{d,e},\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}} - 1)}{(c_{z_{d,e},\bullet,\bullet,\gamma_s} + V\eta) - 1} \\ t_{d,e} \in S: & \dfrac{(c^{-(d,e,\gamma_s)}_{z_{d,e},d,\bullet,\bullet} + \alpha_{z_{d,e}}) \times (c^{-(d,e,\gamma_s)}_{z_{d,e},\bullet,s_{d,e},\gamma_s} + \eta_{s_{d,e}})}{(c^{-(d,e,\gamma_s)}_{z_{d,e},\bullet,\bullet,\gamma_s} + V\eta)} \\ & \times \dfrac{(c_{z_{d,e},\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}} - 1)}{(c_{z_{d,e},\bullet,\bullet,\gamma_w} + V\beta) - 1} \end{cases} \tag{5.3}$$

Given a particular z_i, the probility of word w_i is conditionally independent from

Algorithm 5 Super-Word LDA Algorithm

Input: α, β, η, K, Document Corpus
Output: Z
Detect super-words for each document (i.e. generate image annotations)
Smooth corpus word data, discard super-word tags without a matching text-word
Initialize the z matrix with random topics (1...K)
Calculate initial values for $c_{k,d,\bullet,\bullet}$, $c_{k,\bullet,w,\gamma_w}$, $c_{k,\bullet,s,\gamma_s}$, $c_{k,\bullet,\bullet,\gamma_w}$, and $c_{k,\bullet,\bullet,\gamma_s}$
for $Iter = 1 \rightarrow MaxIter$ **do**
  for all word/super-word tokens in the corpus **do**
    if token is a word $w_{d,e}$ **then**
      Exclude token from $c_{k_{d,e},d,\bullet,\bullet}$, $c_{k_{d,e},\bullet,w,\gamma_w}$, $c_{k_{d,e},\bullet,\bullet,\gamma_w}$
      Sample new topic $k_{d,e}$ using Equation 5.3
      Insert token in $c_{k_{d,e},d,\bullet,\bullet}$, $c_{k_{d,e},\bullet,w,\gamma_w}$, $c_{k_{d,e},\bullet,\bullet,\gamma_w}$
    else if token is a super-word $s_{d,e}$ **then**
      Exclude token from $c_{k_{d,e},d,\bullet,\bullet}$, $c_{k_{d,e},\bullet,s,\gamma_s}$, $c_{k_{d,e},\bullet,\bullet,\gamma_s}$
      Sample new topic $k_{d,e}$ using Equation 5.3
      Insert token in $c_{k_{d,e},d,\bullet,\bullet}$, $c_{k_{d,e},\bullet,s,\gamma_s}$, $c_{k_{d,e},\bullet,\bullet,\gamma_s}$
    end if
  end for
end for

the super-words. This model solves for the set of z by sampling an unknown token $t_{d,e}$ from document d and its associated topic $k_{d,e}$. The probability $p(z_i|w, s, \theta_m)$ is calculated based on the proportionality in Equation 5.3 and the sample used to adjust the model parameters. To properly utilize the Gibbs sampler, the influence of $k_{d,e}$ must be removed from the applicable multinomial distributions, requiring some prior knowledge whether $t_{d,e}$ is a word or super-word.

The SWLDA algorithm is defined in Algorithm 5. It assumes words and super-words are generated by a common vocabulary, permitting shared probabilistic influence when attempting to measure model parameters. The assumption is that this shared vocabulary will improve latent topic models derived using this method over those that only model text [14] or model text and image regions with independent topics [13] [78]. Considering the success of the multinomial Dirichlet distribution in modeling document generative models [14] [48] [114] [78], it seems reasonable to

assume images also generate words fitting a multinomial Dirichlet distribution. By producing super-words from images, the SWLDA algorithm leverages word probability when sampling super-word probability and vice-versa. The model behavior is adjusted by modifying the Dirichlet Priors α, β and η.

While training the model, the fixed Dirichlet priors α, β, and η will emphasize or de-emphasize each term. Increasing α decreases the impact that topics across that document have on the overall topic probability $p(w)$. An α near zero provides the greatest impact for topics within documents, while a τ value of one emphasizes super-word probability $p(s|z)$. To increase the impact of the super-topic on overall topic assignment, one would increase α and β to de-emphasize the word frequency $p(w|z)$ and document frequency probabilities $p(w)$.

The SWLDA model is compared against Latent Semantic Analysis (LSA), LDA, Blob Multimodal Latent Dirichlet Allocation (BMMLDA) and Weighted Term Latent Dirichlet Allocation (WTLDA) using ten-fold cross-validation. LSA and LDA provide a baseline performance measure while BMMLDA (section 5.2.3) is used to test co-clustering against raw image blobs. WTLDA tests the efficacy of sampling the words at a higher rate, then uses the biased LDA model to query by unknown documents. All models were tested against the same data sets and attempts were made to choose parameters that provided good results for each model.

5.1.1 Super-Word Generation.

Super-words are generated using the automated image annotation technique from Chapter IV. WordNet [34] is used to take words surrounding an image and convert them to synonymous sets (synsets). A synset is a lexical grouping of words defined by meaning placed into an ontological hierarchy. This helps distinguish between the vehicle 'plane' and the tool 'plane' when annotating an image. The top synsets are placed

into the set of super-words along with their synonyms. Additionally, each synset has a hyperonymys that defines the broader concept of the synset and hyponymys that define specific instances of the synset. For instance, the earlier example 'plane' has a hypernym of 'heavier-than-air craft'. It includes the hyponymys 'airliner', 'fighter', 'bomber', and 'jet', among others. The hypernymys and hyponymys one level up and down are also inserted into the super-word set, along with each of their synonyms. This helps ensure a variety of words to leverage co-occurrence based on conceptual meaning.

Since the method outlined in Chapter IV is not completely accurate, there will likely be super-words that do not apply to a document's image. To help eliminate these mismatched super-words, those that do not map to a word within the corpus are discarded. This has the added benefit of smoothing the data to improve latent topic extraction.

For the experiment, SWLDA generated super-words using the techniques and parameters defined in [90]. Dirichlet priors of $\alpha = 0.7$, $\beta = 1.01$ and $\eta = 1.1$ produced the highest F-measure. The SWLDA method required 1,000 iterations to converge on an effective model.

5.1.2 Word Generation.

Words are drawn from the document corpus and pruned. Stop words are discarded and low frequency words eliminated. Comparison testing found pruning words that appeared in the corpus fewer than ten times produced good results. Once stop-words and low frequency words were eliminated, some documents contained few remaining words. Documents with fewer than three remaining words were pruned. Comparison testing demonstrated few gains were made by stemming words to their root forms.

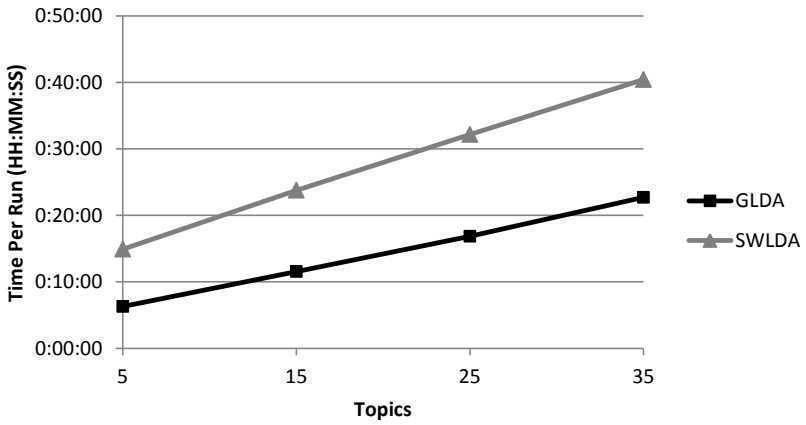


Figure 5.2: Processing Time, GLDA versus SWLDA.

5.1.3 Token Generation.

Words and super-word order are randomized prior to the first iteration. The algorithm then chooses either a word or super-word from the randomized list based on $p_w = \frac{c_w}{c_w+c_s}$ where p_w signifies the probability that a word is chosen, c_w represents the number of words remaining that have not been chosen this iteration, and c_s the number of super-words not yet chosen. Comparison testing demonstrated that, while random word selection requires greater iterations to converge, it results in clustering improvement over using natural word order.

5.1.4 Model Complexity.

The complexity for our implementation of the Gibbs Sampling LDA is $O(WKI)$ where W is the number of words sampled from the corpus, K is the number of topics, and I is the number of iterations to train the model. The additional complexity from adding the super-words changes to $O((W + S)KI)$ where S is the number of super-words sampled. This is not counting the time to generate the super-words from the images.

Experimental verification of this complexity used a laptop with an 2.00GHz Intel

Core i7 CPU using a single core. The actual average processing time for the INitiative for the Evaluation of XML Retrieval (INEX) 2007 test data used in Stage II is calculated for SWLDA and Gibbs LDA. The results in Figure 5.2 match the complexity listed above since processing time should increase linearly with an increase in topics K. The smaller number of additional super-words S causes SWLDA to increase at a faster rate than Gibbs LDA with an increase in K.

5.2 Comparison Models

The SWLDA model is compared against the LSA, LDA, BMMLDA, and the WTLDA models, all described in this section. Using ten-fold cross-validation, each model is tested against the same data using tuned parameters to approximate optimal performance.

Models were tested with a query by document approach [133]. This method uses a previously unseen document, or set of words, to retrieve similar documents from the corpus. While simple queries use keyword searches to return relevant results, they frequently fail to identify synonyms of those keywords. Latent topics often improve query by document since users do not have to match words in the document with keywords exactly. Many papers test latent topic models using the predictive ability of a model, measured as perplexity [14, 110]. Actual retrieval precision and recall, however, directly test query retrieval performance [133, 6]. Only using precision, however, ignores recall, or the number of in-category documents retrieved over all possible in-category documents available (see Equation 2.23). F-measure provides the harmonic mean of precision and recall. Illustrated in Equation 2.24, it is often weighted to emphasize one over the other if necessary.

5.2.1 Latent Semantic Analysis.

As mentioned in Chapter 2, LSA [26] reduces a sparse word occurrence matrix using Singular Value Decomposition (SVD). This results in three matrices, $P = \hat{H}\hat{\Sigma}\hat{V}^t$ where $\hat{H}$ and $\hat{V}^t$ are the left and right singular orthonormal matrices respectively, and $\hat{\Sigma}$ represents the diagonal matrix of singular values. The left and right matrices represent topic and document relationships via a set of scalar values in $\hat{\Sigma}$.

For this test, each matrix vector is sorted by the diagonal matrix $\hat{\Sigma}$ and the most N representative vectors kept while the rest are discarded. Each document in the "unknown" set is compared by first transforming its word occurrence matrix by multiplying $d * \hat{H}\hat{\Sigma}^{-1}$ and comparing the angle between the two vectors. The challenge is in selecting the proper number of vectors N to adequately represent the space while minimizing complexity. Deerwester, et al. [26] suggests selecting the number that "yields good retrieval performance". This research found $N = 200$ worked well. The number of documents returned is user-dependent. To simplify comparison, the number of documents returned is based on the number of topics generated by the other models, or the approximate average size of a latent topic. For instance, if LDA is tested using 15 topics, LSA will return $N/15$ documents where N is the total number of documents in the training set. While this provides a good baseline, these results may be optimized using various clustering methods or other distance measures in future research. LSA is one of the oldest models in the list and provides a baseline measure for all other models.

5.2.2 Gibbs Latent Dirichlet Allocation.

Standard Gibbs sampling LDA [48] provides a pure generative model to compare against SWLDA. Hyper-parameter selection is a complex task with a number of different approaches. Symmetric Dirichlet priors are typically the right answer [124],

though certain instances benefit from asymmetric priors. Increasing β with a symmetric prior de-emphasizes the influence of word topics in relation to document topics while increasing α does the opposite. Griffith and Steyvers [48] suggest $\alpha = 50/T$ and $\beta = 0.1$ where T is the number of topics. Initial comparison testing using the U.S. Trademark and Patent Database from 2011 [2] found best results achieved using $\alpha = 0.4$ and $\beta = 0.5$, though changes in the hyper-parameters had little effect on overall performance. Initial testing using the INEX 2007 data set appeared to contradict the above settings, performing better using a low alpha of $\alpha = 0.05$ and a high beta. Unfortunately, this decreases word probability influence to the point where all documents cluster into the largest four or five topics, eliminating topical discrimination for the small topics. In addition, since document/topic probability has such a disproportionate influence over the model, all but LSA below would produce essentially the same results as LDA. In addition, F-measure becomes biased towards those large topics and provides misleading results. This justified limiting parameters to $\alpha = 0.05$ and $\beta = 1.0$, then keeping them as consistent as possible across the different models. Model convergence was attained by increasing the number of iterations to 700 for this data set. Larger data sets, on average, require more iterations to converge while smaller data sets tend to converge earlier.

5.2.3 Blob Multimodal Latent Dirichlet Allocation.

The SWLDA model takes advantage of information within the embedded images to improve latent topic extraction. Others previously discussed [7] have done something similar by directly using image blobs or blob to token translations. A slightly different model was tested for two reasons. First, those that modeled blobs using topic-based Gaussian distributions failed during comparison testing, likely due to the rough landscape of the data domain. Using eleven topics, the Gaussian models

merged into two primary topics with the other topic Gaussians representing no blobs. Corr-LDA was also considered, but as it uses Gaussian distributions, it would likely encounter the same issues as GM-LDA. Second, the model proposed by [7] and others is primary designed to automatically annotate images using captions where this dissertation attempts to improve document query performance. As the number of words used as captions is increased by extracting surrounding paragraphs, the complexity of the model in [7] increases significantly.

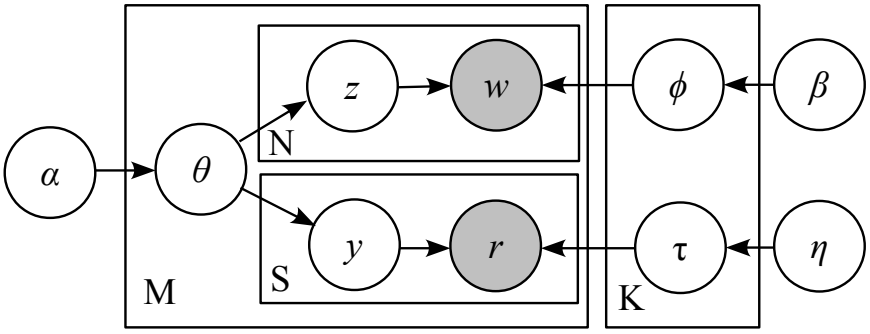


Figure 5.3: Blob Multimodal Latent Dirichlet Allocation Plate Diagram.

To test the efficacy of using blobs to improve latent topics, this research simplifies the model in [7], resulting in the Blob Multimodal LDA (BMMLDA) model in Figure 5.3. Efficient Graph-Based Segmentation with Mean Thresholding [35] [90] segments out blobs in the document and k-means clustering is used to assign similar blobs the same label, symbolized by r. Blob topics, given as y, are sampled independent of words. The multinomial distribution τ represents blob topic probability which derives its parameters from the Dirichlet distribution with prior η.

BMMLDA had the highest F-measure with the largest thirty image blobs from each image. Efficient Graph-Based Segmentation with Mean Thresholding [35] [90] extracts the blobs and k-means clustering assigns labels using 100 clusters. Comparison testing indicated almost no performance gain above 100 clusters. Each blob is assigned a cluster label and Gibbs sampling is performed until model convergence

occurs around 1,000 iterations. This process uses the same parameters as LDA with equal emphasis to word and blob co-occurrence probabilities, or $\alpha = 0.05$, $\beta = 1.0$, and $\eta = 1.0$.

5.2.4 Weighted Term LDA.

The overarching assumption is that images hold topical relevance at or greater than the surrounding words. The BMMLDA algorithm described above uses image features directly. The Weighted Term LDA algorithm utilizes the super-words from Stage I to improve the latent topic models. Following SWLDA [90], this work extracts words surrounding the image, including the caption and surrounding paragraphs. Because many words, such as 'plane', have multiple meanings, this method compares the query image against a set of images in the ImageNet database containing aircraft and a set of images containing tools used to shave wood. The image set with the most similar features is judged as more likely to apply to the query image. Once a list of likely annotations is generated, synonyms are produced using WordNet and the final list is used as higher weighted terms under the assumption that they have higher relevance. The weighted terms are then sampled at a greater rate than standard words during model training, and is the basis behind the WTLDA model.

Comparison testing of WTLDA found sampling super-words thirty times produced one of the higher F-Measures. It was assumed that all thirty samples would receive the same topic assignment. Regular words are sampled a single time each and the model used the same Dirichlet priors ($\alpha = 0.05$, $\beta = 1.0$) and number of training iterations (1000) as LDA.

5.3 Evaluation

Testing of all algorithms uses the same twelve categories from the 2007 INEX data set [28] detailed by Table 4.2. No attempt was made to balance category sizes since real-world data will likely include unbalanced categories. Additionally, this tests the tendency of large topics in Latent Dirichlet Allocation to absorb smaller topics. Ideally, the super-words should provide additional probabilistic leverage towards those smaller topics. Finally, not every image will generate super-words since graphs and clip-art are discarded.

Ten-fold cross-validation is performed on the data, training the model with a randomly-chosen 90% of the documents, then testing retrieval using the remaining 10%. Each of the "unknown" documents are selected and iterated into the model. Words and super-words from the unknown document are used to adjust the model's multinomial parameters temporarily. The dominant topic for each document is calculated based on the topic most represented by the document's words and super-words. From there, precision, recall, and F-measure are calculated based on the documents retrieved for that topic.

5.4 Results and Discussion

The top three annotations from Stage I were incorporated into Stage II as super-words. Each annotation was expanded into its synonyms, hypernyms, and hyponyms and each of those became super-words. Those super-words that did not have an associated word were discarded. Tests were conducted using a range of 3 to 100 topics.

Figures 5.4, 5.5, and 5.6 show that all of the algorithms outperformed the baseline LSA with precision, recall, and F-measure. Of models that incorporated image information, the two that incorporated super-words rather than image blobs, WTLDA

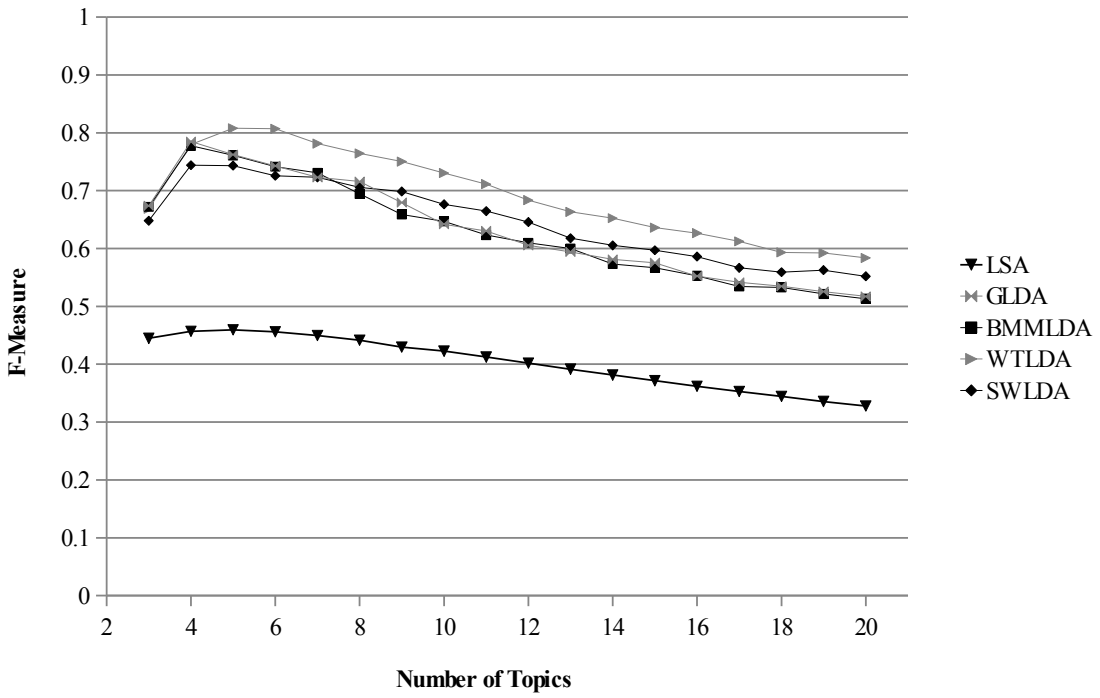


Figure 5.4: F-Measure Comparison, 5 - 100 Topics.

and SWLDA, had higher recall and overall F-measure than all other models when topics were greater than eight. BMMLDA offered little noticeable improvement over Gibbs LDA using the image blob features, though did not decrease much either. In all graphs, an increase in recall often came at the expense of a smaller decrease in precision. The highest F-measure scores often were the lowest in precision, with the exception of LSA which scored low on all measures.

LSA has a gradually increasing precision as the number of documents returned in a query is decreased. This is due to a difference in how LSA tests unknown documents and highlights why topic number is important. Latent topic models apply overlapping topics to the documents and whole topics are typically returned for a query. LSA does, however, calculate vector angles between two documents and can be sorted from least to greatest. As the number of documents returned is shrunk, the remaining documents become the ones most likely to be similar.

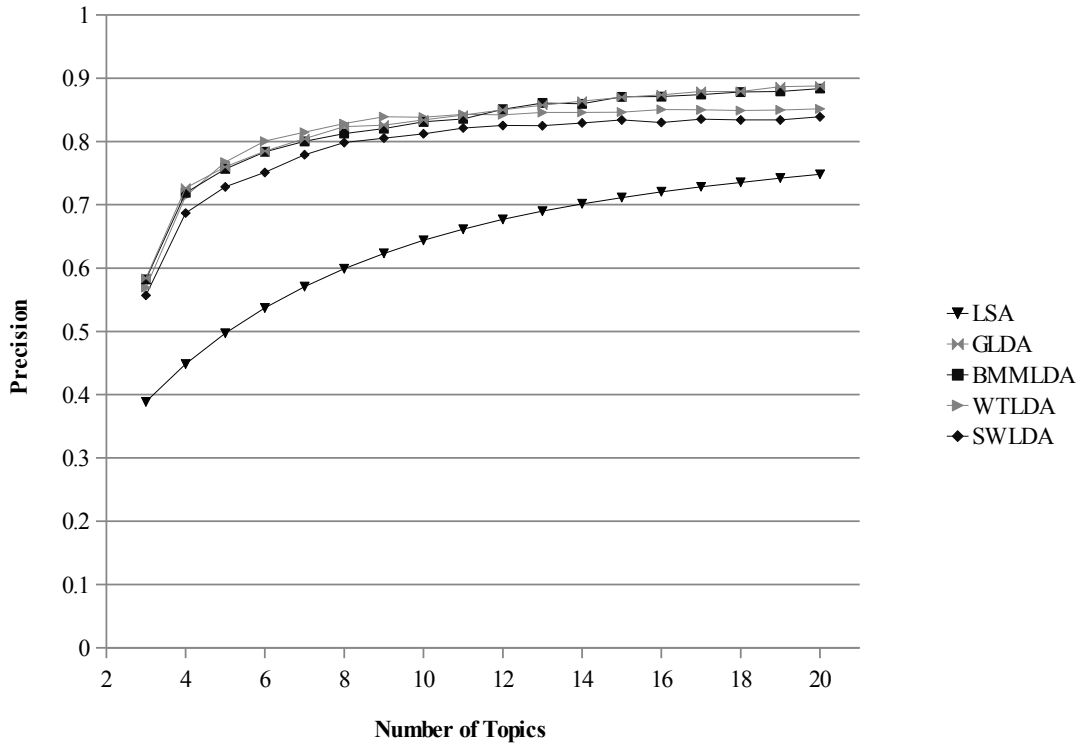


Figure 5.5: Precision Comparison, 5 - 100 Topics.

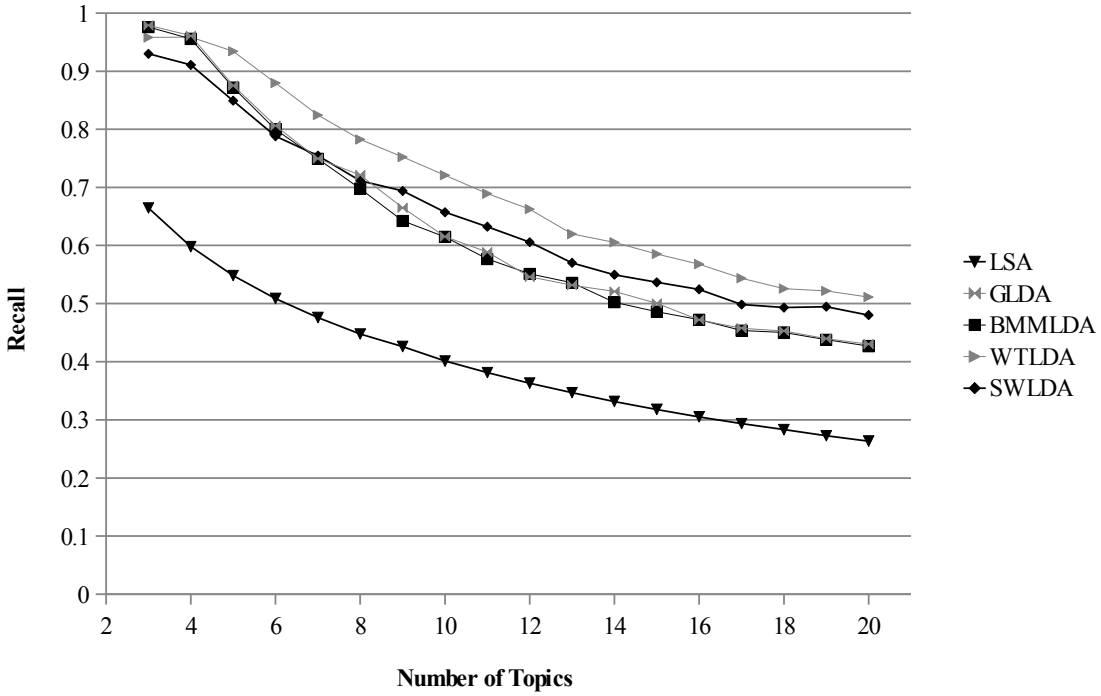


Figure 5.6: Recall Comparison, 5 - 100 Topics.

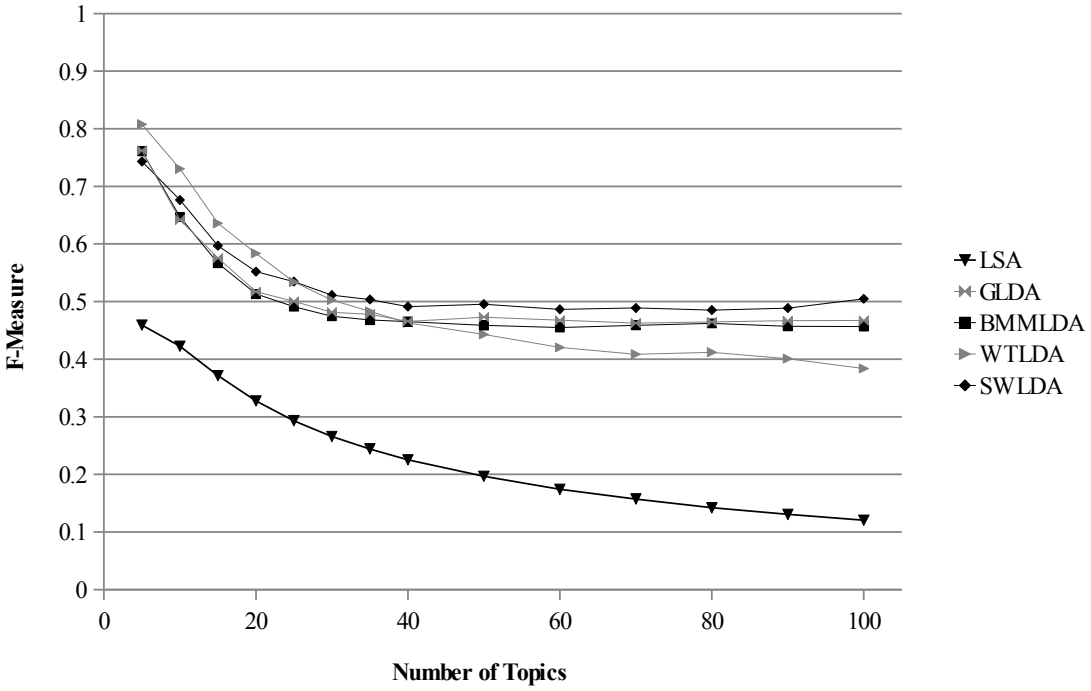


Figure 5.7: F-Measure Comparison, 3 - 20 Topics.

The same phenomena applies to the topical models. The document corpus includes a number of documents that are largely assigned to a topic while some include few topical words or adhere more to a topic that is not well represented by the data set. By increasing the number of topics, topical granularity is improved and the model is better able to match an unknown document. For this reason, precision rises as topic numbers grow. Of course, recall suffers as the entire category can no longer fit into a topic. Performance gains are no longer being made in precision after about twenty topics while recall continues to suffer as topic numbers are increased. Decreasing topics, however, would hide many of the over 6,000 other manually annotated category labels within the document corpus.

As manually generated document categories exist for this data set, query performance can be used to determine the optimal number of topics. In this case, the statistics stop decreasing at around forty topics, indicating the models are unable to

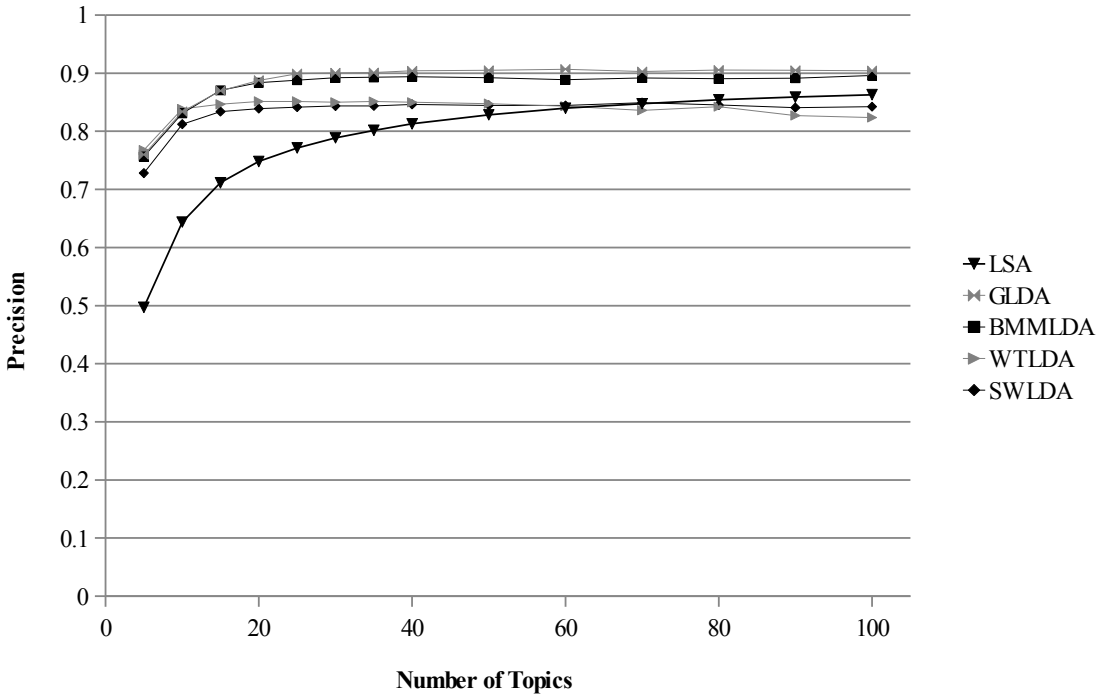


Figure 5.8: Precision Comparison, 3 - 20 Topics.

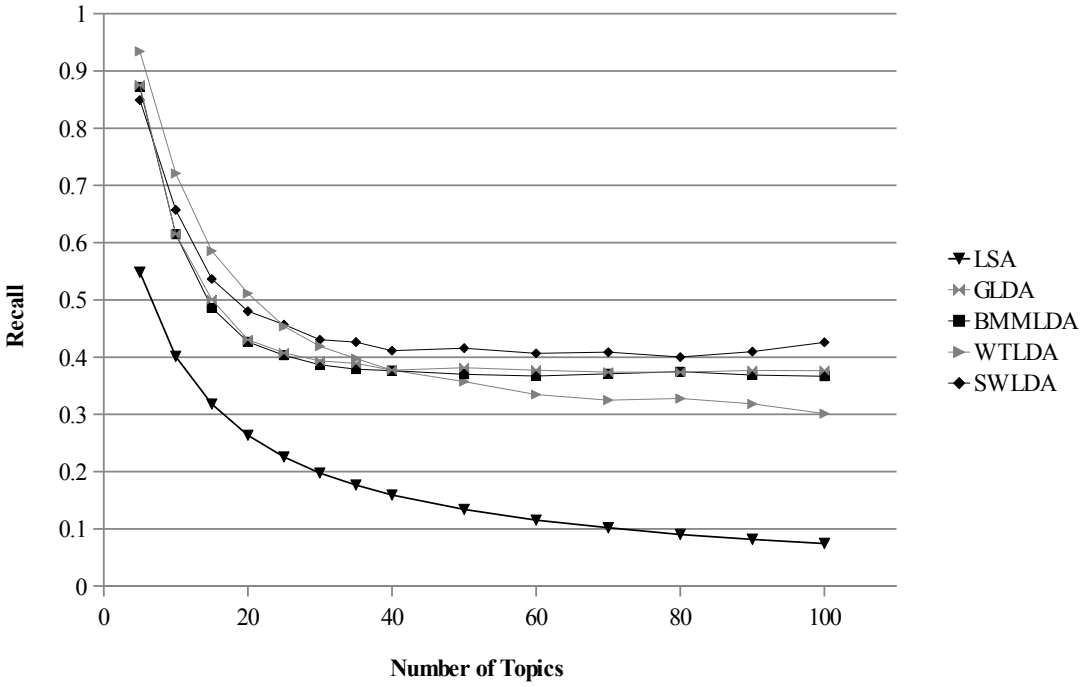


Figure 5.9: Recall Comparison, 3 - 20 Topics.

extract more latent topics beyond that number. When few topics are used, WTLDA has higher precision, recall, and F-measure, supporting the assumption that super-words improve query on latent topic models. As the number of topics are increased, however, it becomes unstable and query performance decreases. SWLDA needs at least ten topics to score higher than Gibbs LDA, as indicated in greater detail by Figures 5.7, 5.8, and 5.9. From that point, SWLDA maintains a consistently high F-measure and recall than all but WTLDA. It overtakes WTLDA at around 25 topics and steadies out in performance at around forty topics, maintaining the highest F-measure score. This indicates that, while sampling super-words at a higher rate with WTLDA provides a high F-measure when the number of topics are small, representing words and super-words using a multimodal approach provides greater stability. Since the optimal number of topics for a data set is typically not known without extensive testing, SWLDA offers the best opportunity for producing a high F-measure.

Table 5.1: SWLDA Clustering Matrix Using 11 Topics.

Topic	1	2	3	4	5	6	7	8	9	10	11
Dogs	0	0	1	621	2	7	7	64	16	2	7
Mountains	0	0	8	2	1569	9	120	7	322	10	2
Aircraft	93	1128	18	1	3	344	2	9	2	3	160
WWII Ships	0	1	709	0	0	1	0	0	1	0	0
Armored Vehicles	0	58	1	0	0	3	62	0	1	0	0
Flowers	0	0	0	0	1	0	1	10	1	138	0
Skyscrapers	0	0	2	2	207	2	2	1	3	2	7
Elephants	0	0	4	5	1	4	0	3	21	1	1
Sailboats	1	2	80	2	4	10	0	15	5	1	1
Vegetables	1	0	0	3	1	1	0	3	0	277	0

Table 5.1 provides an example of a SWLDA clustering matrix. It shows how each category clusters within the topics and provides insight into the advantages of SWLDA over Gibbs LDA (Table 5.2). While the improvements to recall are visible when comparing Tables 5.1 and 5.2, the loss of precision is also apparent. SWLDA tends to produce a small scattering of a category across a number of topics. While

Table 5.2: Gibbs LDA Clustering Matrix Using 11 Topics.

Topic	1	2	3	4	5	6	7	8	9	10	11
Dogs	15	0	1	0	0	2	227	474	0	5	1
Mountains	1	1	1073	1	0	880	63	3	6	3	0
Aircraft	0	5	0	382	108	0	34	0	26	353	877
WWII Ships	0	364	0	0	0	0	1	0	342	1	0
Armored Vehicles	0	0	0	125	0	0	2	0	2	0	0
Flowers	132	0	0	0	0	0	11	0	0	0	0
Skyscrapers	0	0	0	0	0	220	4	0	0	1	0
Elephants	4	0	0	0	0	0	35	0	0	1	0
Sailboats	0	5	0	0	0	0	18	0	93	3	2
Vegetables	284	0	1	0	0	0	10	0	0	0	0

it includes a relatively small proportion of each category, users will need to consider whether the loss of precision is worth the gains in recall.

Table 5.3: SWLDA Clustering Matrix Using 5 Topics.

Topic	1	2	3	4	5
Dogs	3	6	5	4	709
Mountains	4	1986	13	21	10
Aircraft	1724	6	31	5	14
WWII Ships	2	2	694	0	0
Armored Vehicles	119	1	5	0	1
Flowers	0	2	0	143	0
Skyscrapers	2	31	3	188	5
Elephants	0	26	0	2	10
Sailboats	16	9	90	3	6
Vegetables	1	1	0	295	4

Tables 5.3 and 5.4 help illustrate how Gibbs LDA performs better at lower topic levels. The four categories with the largest number of documents are 'Mountains', 'Aircraft', 'Dogs', and 'WWII Ships'. When the topic number equals '5', Gibbs LDA draws those four categories, each into their own topic, with one topic left over for the rest of the data. The low number of topics helps prevent the usual topic split that Gibbs LDA does to the four largest categories. Yet, the problem with low frequency scattering still exists with SWLDA. Hence, at this point, Gibbs LDA has

Table 5.4: Gibbs LDA Clustering Matrix Using 5 Topics.

Topic	1	2	3	4	5
Dogs	710	2	16	2	0
Mountains	11	1	36	1976	4
Aircraft	23	1742	7	1	15
WWII Ships	0	3	0	1	695
Armored Vehicles	0	131	0	0	2
Flowers	1	0	143	0	0
Skyscrapers	1	0	238	0	0
Elephants	10	0	25	0	0
Sailboats	1	5	24	2	85
Vegetables	10	0	278	1	0

better precision, recall, and subsequently F-measure. It is, however, clustering six out of the ten total categories into one topic and losing the ability to discriminate. If the categories included an equal number of documents, the precision value would illustrate this loss of performance.

5.4.1 Model Comparison.

Table 5.5 provides a summary comparison between the various topic models tested in this dissertation. LSA was the baseline and certainly the fastest, but had the lowest F-measure score. LDA was faster than any of other methods beyond LSA since it had the lowest complexity and had no significant performance difference between BMMLDA, though it did not incorporate images and, therefore, would not be able to make any inferences regarding images. WTLDA had high performance with lower numbers of topics, but lacked stability so performance fell off as the number of topics was increased. The other models maintained fairly consistent results as topics were adjusted.

Table 5.5: Latent Topic Model Comparison

Model Name	F-Measure	Speed	Supports Images	Stability
LSA	L	H	N	H
LDA	M	H	N	H
BMMLDA	M	M	Y	H
WTLDA	H	M	Y	L
SWLDA	H	L	Y	H

5.5 Conclusion

The SWLDA process and testing in this section supports the hypothesis that incorporating highly-influential super-words into the LDA model can improve model prediction over similar models for unknown documents. Once the hidden parameters of the model are estimated, posterior probabilities can be used to draw inferences. The super-word posterior probabilities indicate which super-words are more likely given a particular document. This information can be forwarded back to the automated image annotation algorithm to improve results from Stage I. This technique is further described and tested in Stage III, detailed in the following chapter.

VI. Stage III: Automated Image Annotation Refinement

Stage I produced image annotations based on the words surrounding an image and the features of that image. While it demonstrated improved results over generalized image annotators, there were still a number of poor annotations. This research hypothesizes that the poor annotations will also be topically irrelevant to the document. This means that posterior probability from the Stage II Super-Word Latent Dirichlet Allocation (SWLDA) model in the previous chapter (indicated in Figure 6.1) can be used to select topically relevant super-words over out-of-topic words. This chapter describes stage III, or the automated image annotation refinement stage, of the Automated Image Annotation and Latent Topic model. It first discusses the methodology and experimental design, then summarizes and analyzes the results.

6.1 Methodology and Experimental Design

Posterior probabilities can align image annotations more closely with document topics. The risk is that this could create a homogeneous set of annotations and decrease word specificity. Ideally, the method chosen to combine the image and text posterior probabilities would maintain the diversity in both topic probabilities, local word context, and image features.

This section first discusses several options for posterior probabilities and indicates which was chosen. Next, it discusses potential means for combining the Earth Mover's Distance (EMD) from Stage I and the SWLDA posterior probability from Stage II, along with the comparison testing performed on each. This is followed by the experimental set up and results that show improved annotations

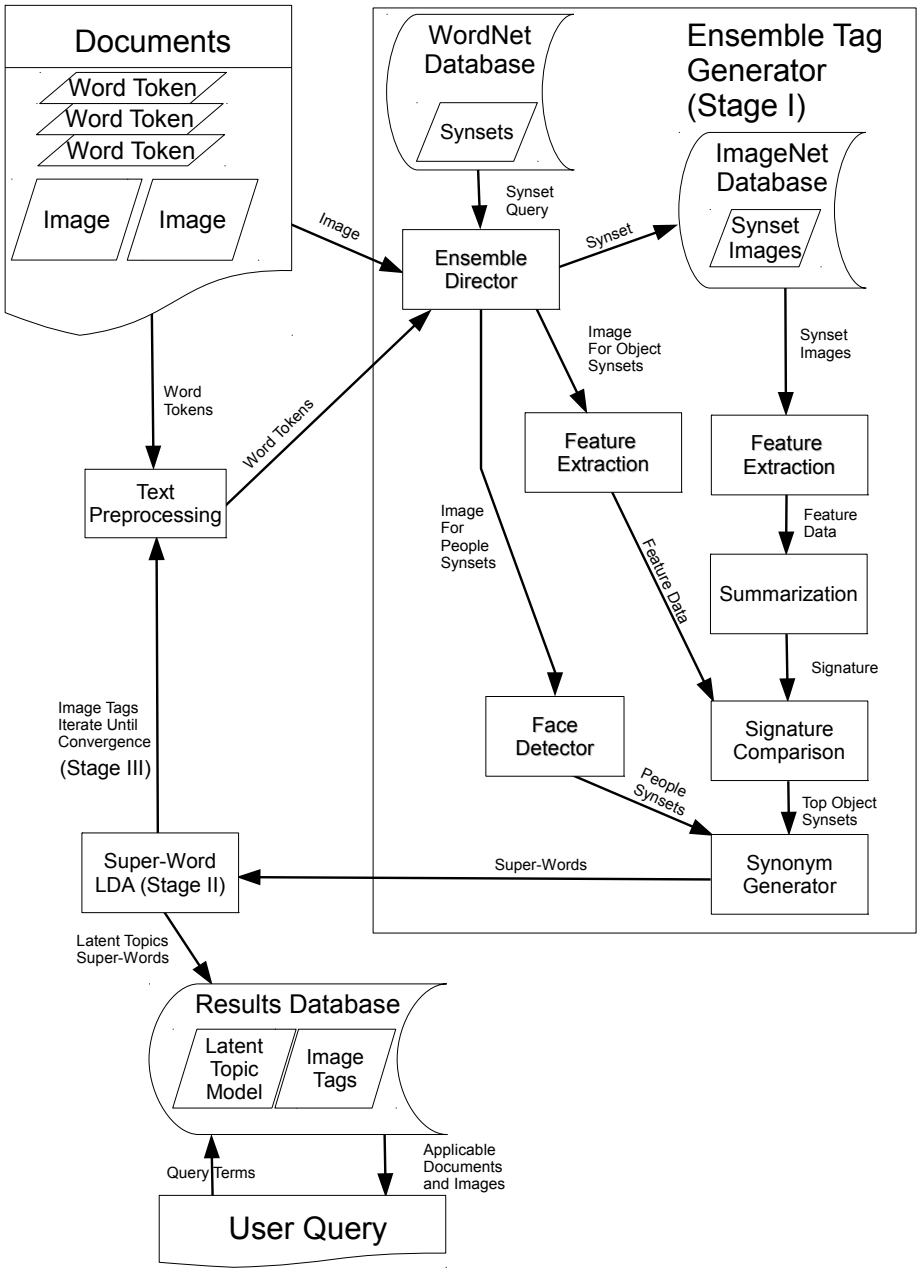


Figure 6.1: Automated Image Annotation and Latent Topic Iterative Model.

6.1.1 Options for Calculating Posterior Probability.

There is not a direct probabilistic relationship between data produced by the automated image annotation techniques in Stage I and posterior probabilities in Stage

II. This section presents three options for combining these values.

6.1.1.1 Posterior Probability Option 1: Word Probability.

The first option looks at word posterior probability independent from super-word probability, or $p(w_i|\theta_d, \phi_i)$ where w_i is the word indexed by i and θ_d is the multinomial probability of the document/topic probability for document d. The multinomial probability ϕ_i provides the word/topic probability indexed or word i.

6.1.1.2 Posterior Probability Option 2: Super-Word Probability.

The second option uses solely super-word probability independent from word probability. This finds $p(s_i|\theta_d, \eta_i)$ where s_i is the super-word indexed by i and η_i is that super-word's word/topic multinomial probability.

6.1.1.3 Posterior Probability Option 3: Combined Probability.

Finally, we can use a hybrid weighted combination of the two probabilities with $p_c = w_1 * p(w_i|\theta_d, \theta_i) + w_2 * p(s_i|\theta_d, \eta_i)$.

$$p(s_i|\theta_d, \tau_i) = \frac{\sum_{t=1}^{T} p(z_{d,t})p(s_{i,t})}{\sum_{w=1}^{S}\sum_{t=1}^{Z} p(z_{d,t})p(s_{w,t})} \tag{6.1}$$

The posterior probability of the super-word given a particular document is defined by Equation 6.1. The $p(z_{d,t})$ is the probability that a document has a particular topic, while $p(s_{i,t})$ is the probability that a super-word is in a particular topic. The denominator is the normalizing summation. Comparison testing demonstrated the super-word probability provided the best image annotation improvements while word probability and a weighted combined probability performed worse.

6.1.2 Options for Combining Posterior Probability with Earth Mover's Distance.

While word or super-word posterior probability provides the likelihood that a particular word or super-word will be generated, it does not take into account image features. An algorithm that is able to incorporate both posterior probability and image feature information may be able to build improved image annotations. Combining posterior probability and image features is challenging. This section lists four options for combining posterior probability with image features.

6.1.2.1 Option 1: SWLDA Threshold, EMD order.

The first option takes the top ten words based on super-word posterior probability, then ranks those words using descending EMD. This method has the benefit of pruning the search space even further by eliminating unlikely words based on the document topic, then using their image features to select the best words. This option tended to produce poor results in comparison testing, likely since initial evidence indicates super-word posterior probability does a better job after the first iteration than EMD in annotating images.

6.1.2.2 Option 2: Average Merge.

This option averages the two lists together, producing a combined list. A word in position '1' for one list and position '5' for another would average to position '3'. This is a very simple solution though it makes the assumption that each list is equally accurate. If we determine one list tends to be more accurate than the other, we can weight the list. Therefore, the position of word i in the overall list, or p_i^o is based on a weighted average of the position of the word in both lists. $p_i^o = \frac{p_i^l * w_1 + p_i^a * w_2}{w_1 + w_2}$ where p_i^a is the position on the image annotation list and p_i^l is the position on the latent

topic probability list. w_1 and w_2 are user-defined weights. This option was one of the better performing options with a typically higher relevancy score to automated image annotations. It was not, however, the best algorithm from a relevancy standpoint and it required tuning the weights for ideal performance.

6.1.2.3 Option 3: Normalized Weights Based On Posterior Probability and Earth Mover's Distance.

This option uses weights based on normalized probability and normalized EMD to construct an overall ranking. This accounts for words with very high probability compared to their peers or very low EMD. In this case the ranking r_i is given by

$$r_i = w_1 \frac{p(v_i)}{\sum_i p(v_i)} \times w_2 \frac{MAX_{EMD} - EMD_i}{\sum_i MAX_{EMD} - EMD_i} \tag{6.2}$$

where w_1 and w_2 are user-defined weights, $p(v_i)$ is the probability drawn from the latent topic model for word v_i, and EMD_i is the EMD for word i. MAX_{EMD} represents the maximum EMD value of all applicable words and is used to convert EMD into a measure where smaller values signifies less of a link between word and image. The advantage of this method is that words with extremely low applicability will be penalized appropriately while those with high applicability in one list may not be penalized as much by mediocre applicability in the other list. Comparison testing demonstrated this method offered little improvement over option 2.

6.1.2.4 Option 4: Rank Solely by Super-Word Posterior Probability.

This rank-orders the words solely by $p(s_i|\theta_d, \eta_i)$. The initial super-words are derived from the image annotation algorithm which helps prune the space to a subset. The assumption is this subset of words is more likely to produce relevant annota-

tions and, taking these high relevancy annotations, sorting them by posterior probability produces further refined annotations. Comparison testing demonstrated this approach produced the most consistently high results of the four methods.

6.2 Iteration and Annotation Expansion

Ideally, Stage I and Stage II will prune the list to the most relevant words. Higher confidence in the remaining words permits their further expansion to ensure an assortment of highly relevant words. Higher hypernym and lower hyponym levels can be added to the list, processed through the automated image annotator, then pruned using SWLDA. This can be repeated until convergence, though defining convergence is not straightforward. At a certain point, adding more words to the list of possible words fails to produce higher relevance words and creates categorical overlap. Additionally, expanding hypernyms could cause less specific words to push out more specific ones.

Comparison testing demonstrated some benefit to expanding hypernyms/hyponyms to create a larger pool of potentially applicable words. Not expanding for an iteration decreases noise and helps prune off some of the poorer performing words. Potentially, a solution would be to switch off, expanding one iteration and not the next. This research tests both approaches, expanding each iteration and expanding every other iteration.

Of course, as super-words are pruned using this iterative approach, the number of image annotations shrink. These can be expanded by exploring synset hypernyms and hyponyms since their meaning will be similar, yet they may provide more relevant annotations. As comparison testing discovered, expanding hypernym and hyponym results too early can create a large number of inaccurate words and result in an imprecise model. Ideally, once the list of super-words has been pruned to a small

number of highly relevant annotations, they can then be expanded and pruned using an iterative approach.

6.3 Experimental Design

This method is tested using the same INitiative for the Evaluation of XML Retrieval (INEX) 2007 Wikipedia Dataset [28]. Performance is verified using Amazon Mechanical Turk (AMT) scores. Five AMT workers rate the relevancy of a word to an image from one to five, where one indicates not at all relevant and five indicates completely relevant. The average score between the five workers is calculated and used as the standard. Results are generated automatically using the method defined in this dissertation and the results compared against the AMT scores. The AMT score of the top five words selected are averaged to produce a score for that image.

6.3.1 Pruning and Smoothing.

Words are drawn from the corpus and pruned in the same fashion as Stages I and II. The same level of pruning was used for each iteration of the algorithm, drawing the text from the original source documents. This means that the regular document words will not change even though the super-words will be refined with each iteration. Super-words that lack a representation within the regular word set are discarded with the assumption that words not appearing elsewhere in the corpus are unlikely to apply to the image.

Similar to Stage II, words and super-word order are randomized prior to the first iteration of the SWLDA algorithm. The algorithm then samples a word or super-word from the randomized list based on a probability, defined as $p_w = \frac{c_w}{c_w + c_s}$. In this equation, p_w signifies the probability that a word is chosen, c_w represents the number of words remaining that have not been chosen this iteration, and c_s the number of

Table 6.1: Stage III Results - Hypernym/Hyponym Alternating Expansion

	1	2	3	4	5	6
Avg Score - Best Case	1.985	2.110	2.134	2.213	2.215	2.239
Avg Score - Worst Case	1.985	2.096	2.119	2.115	2.059	1.952
# Super-words	30245	6804	99516	6555	55291	6905
Avg Words Per Image	6.53	7.00	3.15	4.58	2.33	4.25
SWLDA F-Measure	0.512	0.504	0.491	0.501	0.491	0.488
Hyper/Hypo Expansion	Y	N	Y	N	Y	N

super-words not yet chosen.

6.4 Results and Discussion

Image annotation refinement used the results from both the automated image annotation in Stage I and the SWLDA posterior probabilities in Stage II to refine image annotation. The super-words from Stage II are rank-ordered for each image based on their SWLDA posterior probabilities and scored as in Stage I. However, there was one main limitation of the data. In Stage I, all results were validated by Amazon Mechanical Turk (AMT) workers. Due to insufficient resources, accurate scoring could only be performed in most of the first three iterations. To compensate for this, two values are calculated. The first extends the average scoring for that image to all "unknown" image/word pair scores. While this is considered the "best-case" scenario, a true "best-case" would be to assume all missing scores are '5'. This is, however, unrealistic so the average score is used. The second row provides the worst-case score, assuming unscored image/word pairs receive a '1'.

Tables 6.1 and 6.2 provide the results of two different methods. The first method expands words into their hypernym/hyponym forms every other iteration. During the non-expansion iterations, words are pruned based on low posterior probability. The second method, with results described in Table 6.2, expands hypernym/hyponym forms each iteration and produces far more super-words, but dilutes the average

Table 6.2: Stage III Results - Hypernym/Hyponym Full Expansion

	1	2	3	4	5
Avg Score - Best Case	1.985	2.144	2.197	2.232	2.296
Avg Score - Worst Case	1.985	1.995	1.887	1.871	1.882
# Super-words	30245	88021	48174	78628	55291
Avg Words Per Image	6.53	7.00	4.65	4.29	3.69
SWLDA F-Measure	0.512	0.506	0.496	0.521	0.500
Hyper/Hypo Expansion	Y	Y	Y	Y	Y

score for that iteration. Both tables use super-word posterior probability, or Option 4 defined in Section 6.1.2.4.

The best-case score demonstrates that, as new words are added through expansion, the worse performing words drop out. This causes the best-case score to climb, though this may be misleading since the "unknown" scores could be unrelated word/image pairs. Assuming that every new word without an AMT score is scored as 1, we see the potential for hypernym/hyponym expansion to result in poor image annotation.

Attempts were made during comparison testing to not expand hypernyms or hyponyms during any iteration. Unfortunately, this did not improve automated image annotation relevancy, but only served to decrease the average number of words per image through constant pruning.

Table 6.2 expands hypernym/hyponyms every iteration and quickly reaches the "best case" score. It does, however, result in a rapid decline of the worst-case score, indicating that a few good annotations are being replaced by potentially bad annotations.

The first iteration of these two tests demonstrates an improvement in image annotation relevancy using SWLDA posterior probability. SWLDA performance is tested to determine if document retrieval, scored by F-measure, improves using the improved image annotations. While it does not get significantly worse, it also does not appear to improve. One hypothesis is that each iteration moves the model away from annota-

Table 6.3: Sample Automated Image Annotation Refinement Results Using Aircraft Image.

First Iteration	AMT Score	Third Iteration	AMT Score
line (cable)	1.0	aircraft	5
aircraft	5.0	craft	4.2
line (production)	1.0	—	—
line (telephone)	1.0	—	—
guppy	1.0	—	—
Average	1.8		4.6

tions derived using image features and more towards corpus word probability without super-word influence.

Table 6.3 provides an example of an image and its super-words. Based on the local context and image features, the words aircraft, three variations of the word line, and a fish variety were chosen. Aircraft was expanded into craft during iteration two, then pruned during iteration three. This process eliminates the unrelated "line" words and the guppy, leaving the two relevant words.

Table 6.4 provides a second illustration, demonstrating the changes made using super-word posterior probability. 'Button' is expanded into its alternate form 'push' and 'switch' while 'bridge' is likely expanded in error from 'crane'. Finally, 'radius' and 'spoke' are pruned off as topically irrelevant.

The final example in Table 6.5 identifies three words from the local context and is based off image features. Annotations 'hound' and 'coonhound' apply directly to the image while 'basset' is close. Notice that 'basset' was rated higher than the perfect

Table 6.4: Sample Automated Image Annotation Refinement Results Using Dog Image.

First Iteration	AMT Score	Third Iteration	AMT Score
radius	1.0	button (push)	3.8
button (fastener)	1.0	push (button)	3.8
button (push)	3.8	switch	4.2
crane	1.0	bridge	1.0
joystick	4.6	—	—
Average Score	2.3		3.2

Table 6.5: Sample Automated Image Annotation Refinement Results Using Joystick Image.

First Iteration	AMT Score	Third Iteration	AMT Score
hound	4.8	basset	4.8
coonhound	3.4	coonhound	3.8
black	1.0	wolfhound	3.8
	—	—	—
	—	—	—
Average	3.1		4.1

fit of 'coonhound', indicating some uncertainty and variability in human annotations due to lack of information on specific dog breeds. After the hypernyms and hyponyms are expanded, then pruned once more, the unrelated word 'black' is pruned off and replaced by a closer word 'wolfhound'. Again, even with humans providing annotations, they scored 'coonhound' the same as 'wolfhound', even though the former looks closer than the latter.

These three examples provide an idea of improvements using posterior probability. While many do not perform as well as these examples, enough improve to increase the overall score. Much of the increase comes from pruning poor words, however, some new words are added.

6.5 Conclusion

This chapter demonstrates that posterior probability from topic modeling (Stage II) can help improve automated image annotation. By pruning topically irrelevant words and expanding topically relevant words, image annotations can be improved with only a small loss in specificity. Also iterating the entire model by feeding image annotations back to Stage I can further improve image annotation.

VII. Part II: Applicability of Latent Dirichlet Allocation to Multi-Disk Search

For a forensics analyst, data stores will likely include a variety of disjointed documents with sometimes topically-irrelevant embedded images. Much of a hard drive may include directories of images without much context. With all that noise, extracting topically relevant data is tricky. Responding to a query for topically-relevant data is tougher.

Part II attempts to address whether the models described in this dissertation can be used by the digital forensics analyst to extract meaningful information from a corpus of real-world digital forensics data. To answer this question, we first calculate statistics on the user-generated data in the Real Data Corpus (RDC) [42] to identify areas suitable for search evaluation. Then we use a series of tests that replicate common digital forensics tasks. The first test uses a set of queries on three different topics. Search using Latent Dirichlet Allocation (LDA) is compared against traditional regular expression search based on the relevancy of returned results. The second test measures the ability of LDA to topically segment a corpus, then separate a selected topic into subtopics. Finally, the third test compares query using LDA to query using regular expression in a topically noisy environment. The section first discusses the methodology and experimental design, then summarizes and analyzes the results.

7.1 Methodology and Experimental Design

The Naval Postgraduate School and University of North Carolina maintain a collection of hard drives called the Real Data Corpus (RDC) [42]. It contains a wide range of disk images extracted from storage devices purchased on the open market and separated by country. The devices range from 8MB to 480GB hard drives and in-

cludes data previously stored on phones, flash cards, USB drives, and multi-partition hard drives. These drives are stored as Encase files, Advanced Forensics Format (AFF) files, or in raw form. The entire corpus used in this testing consisted of 2,435 disks from twenty-five different countries.

The remainder of this section describes the process used, including how images are verified, mounted, and searched. It then discusses how the latent topic models were run against the corpus.

7.1.1 Disk Image Extraction.

Since this research is primarily interested in user-generated data, some processing was required to prune irrelevant data. Disk images consisted of several formats, including the Encase file format, Advanced Forensics Format (AFF), and raw disk images. All steps were accomplished on a Ubuntu Linux system. Encase files were first verified using ewfverify. They were then mounted using ewfmount and any errors in the process documented. AFF files were mounted using 'affuse'. Once the specialized image was mounted, the raw image could be accessed and mounted to a loopback device using 'losetup', then the file system mounted using 'mount'. Raw image files were mounted directly using 'losetup' and 'mount'. File systems were then traversed using Java code.

7.1.2 File Extraction.

Of the 2,435 drives, 920 had user-created files. Exclusion directories had to be identified for some common directories with large numbers of boilerplate files. These included common systems directories, such as C:\WINDOWS and frequently seen applications. Further pruning used a database of SHA1 hashes so duplicates could be identified and purged. All remaining interest files were read into the Java function

MessageDigest in the java.security library and the resulting strings loaded into the MySQL database for comparison against other files. Additionally, only files with English were included to constrain scope. Finally, only files that matched one of a set of extensions were extracted. Text files were read using simple Java input functions while images were loaded using OpenCV. The Java LibreOffice API 4.1.0.0a library was used to load Microsoft or OpenOffice applications and parse the text and images.

Each document is processed using a variety of techniques. Microsoft Word documents have their text and images extracted and stored in a database. Images are exported to the disk and information about the image stored in a database table. Raw text files tended to be either logs, application data, readme files, or e-mails. Since e-mails are the only applicable documents for this research, they were identified using a regular expression search and extracted while all others were discarded. Extensible image file format (exif) information is extracted from image files and stored in the same database. Web browser caches were extracted separately from the rest of the files since those aren't user generated data. Browsers' caches were extracted from Microsoft Windows machines and included Internet Explorer, Netscape and Firefox.

Several tests were accomplished to identify areas promising for image annotation and Super-Word Latent Dirichlet Allocation (SWLDA) experiments. Each image is checked to assess if it looks like a graph using the method defined in Section 4.3.2. Faces are detected using a boosted cascade of Haar-like features [123]. To maintain consistency across image analysis and to improve speed, they are shrunk to 400 pixels wide, then face detection is performed on the image. An initial sampling of the Microsoft Word documents indicated the majority are business documents. Most embedded images are of graphs, diagrams, or close-ups of machine hardware.

Experimental design can not account for issues arising from the fact that the data consists of real-world hard drives from an assortment of disk images. It lacks valida-

tion information, and document categories must be manually assessed to determine clustering effectiveness. Additionally, one hard drive may not contain enough information to extract latent topics and precise image annotation. Using documents from several hard drives, however, joins together data that likely has no connection. If a set of hard drives exist that appear to be from a similar source, such as those from the same company or university, combining the data may prove useful. All these decisions were manually assessed and noted for each test.

7.1.3 Latent Topic Search.

Digital forensics practitioners rely mainly on keyword search to identify useful information [71]. Some best practices do exist when crafting keyword searches [85], however, most rely on experience and intuition to build a keyword list likely to produce relevant query results and limit false positives. Large corpus compound the search problem increasing false positives and expend the available manpower.

In addition to proper keyword selection, regular expression search requires intelligent parameter tuning. Regular expression search benefits from term weighting since some terms often have higher relevance for a topic than others. Ideally, relevant documents will contain a number of keywords with high frequency. For this research, the document list is sorted first by the unique keywords found, then by the keyword score. The score s is calculated as $s = \sum_{i=1}^{K} w_i k_i$ where K is the set of keywords, k_i is the frequency of keyword i and w_i is the weight of keyword i. The 'unknown' target documents are identified and used to compare each algorithm based on its position in the query results.

The LDA tests are designed to be as close to the regular expression search as possible, including using the same keywords where applicable. Once the LDA model parameters have been estimated, they can be used to draw conclusions about new

data. In this research, the LDA topic model is trained using Gibbs sampling, varying iterations based on corpus size. For example, models formed using the Israeli corpus are trained using 1,400 iterations due to its large size while models trained on the Serbian corpus only required 1,000 iterations to converge. LDA fixed prior Dirichlet parameters are $\alpha = 0.05$ and $\beta = 1.0$. Unless otherwise specified, twenty-five topics are used as it provides acceptable granularity in comparison testing while maintaining reasonable algorithmic speed. Equation 7.1 defines the probability that a document d_i relates to the provided set of keywords X. Multinomial parameters ϕ and θ define the document-topic and word-topic probabilities for each topic K. Weights are used to stress certain probabilities over others and are defined as w_j, the weight for keyword x_j.

$$p(d_i|X,\theta,\phi) = \sum_{k=1}^{K} \left(p(d_{i,k}|\phi) \frac{\sum_{j=1}^{X} w_j, p(x_j|\theta)}{\sum_{j=1}^{X} w_j} \right) \tag{7.1}$$

Word documents were chosen over other document types since they had the largest quantity of text user data. Adobe PDF documents tended to be related to user manuals or other documents downloaded from the internet. To avoid confusing the algorithm, these were excluded. Similarly, Microsoft Excel documents and Powerpoint presentations lacked sufficient text to assist in latent topic extraction.

7.2 Real Data Corpus Analysis

Of the 920 disks that had user-created files, the top ten disks with respect to file count held 36% of the total number of files. The mean number of files per image was 977 with a medial of 77 and the largest drive held 109,938 of the interest files. This appears to indicate that most computer users do not have a significant amount of data on their computers, while a handful of users generate a large number of files.

Of the 25 countries with disks in the corpus, Israel had the largest number of user-

Table 7.1: Drive and File Statistics by Country.

Country (code)	Total Drives	Drives Loaded	Drives Failed	Total Interest Files
Israel (il)	297	283	14	146,297
India (in)	672	566	106	90,390
Mexico (mx)	175	175	0	85,479
Palestinian State (ps)	140	126	14	76,566
Serbia (rs)	8	7	1	23,184
Serbia and Montenegro (cs)	16	15	1	16,850
Canadian (ca)	18	11	7	14,173
Singapore (sg)	34	23	11	8579
Turkey (tr)	10	10	0	7302
Panama (pa)	17	15	2	6434
China (cn)	808	808	0	5141
Unknown (nnn)	109	109	0	4072
Egypt (eg)	7	7	0	3734
Ukraine (ua)	57	50	7	3577
Pakistan (pk)	85	82	3	2372
Ghana (gh)	21	20	1	1506
Germany (de)	4	2	2	1444
United Arab Emirates (ae)	39	34	5	624
Japan (jp)	13	13	0	423
Hong Kong (hk)	4	4	0	320
Morocco (ma)	11	10	1	44
Bangladesh (bd)	57	54	3	18
Switzerland (ch)	2	2	0	0
Hungary (hu)	22	2	20	0
Thailand (th)	4	4	0	0
Total	2,630	2,432	198	98,529

generated interest files, as indicated in Table 7.1. Though many files were written in Hebrew, a number were written in English, making it the richest disk set in terms of searchable English data. The Indian collection held a number of files from a variety of disciplines, including political, business, and scientific topics. Though China had the largest number of drives, they held few files in English beyond operating system and application files. Those few Microsoft Word documents that did appear were mostly written in Mandarin. Likewise, the Mexican corpus included a large number of interest files, but most were in Spanish. This dissertation constrains the scope by primarily using English documents, though there is ample data for research in other languages.

7.2.1 File Types.

Figure 7.1 provides file extensions of extracted files by frequency in descending order. Images are a significant portion of the overall file count, with JPEGs as the most common file. A majority of these images were personal photographs taken with digital cameras. Others largely consisted of images downloaded from web pages or unique images that were part of an application. The corpus contains over 350,000 MP3s, predominantly music.

Figure 7.2 lists the file types by country. The preponderance of JPEG files is apparent in most countries, but some differences stand out. For instance, the Palestinian State (ps) has few JPEGs but a large number of TIF files. Israel and Mexico have the largest corpus of Microsoft Word documents, though the Mexican documents are written mostly in Spanish while the Israeli corpus has a number of English documents.

The HTML files listed in Figure 7.1 are those found outside of the browser cache. Those files were extracted too, but are counted separately since they reveal differ-

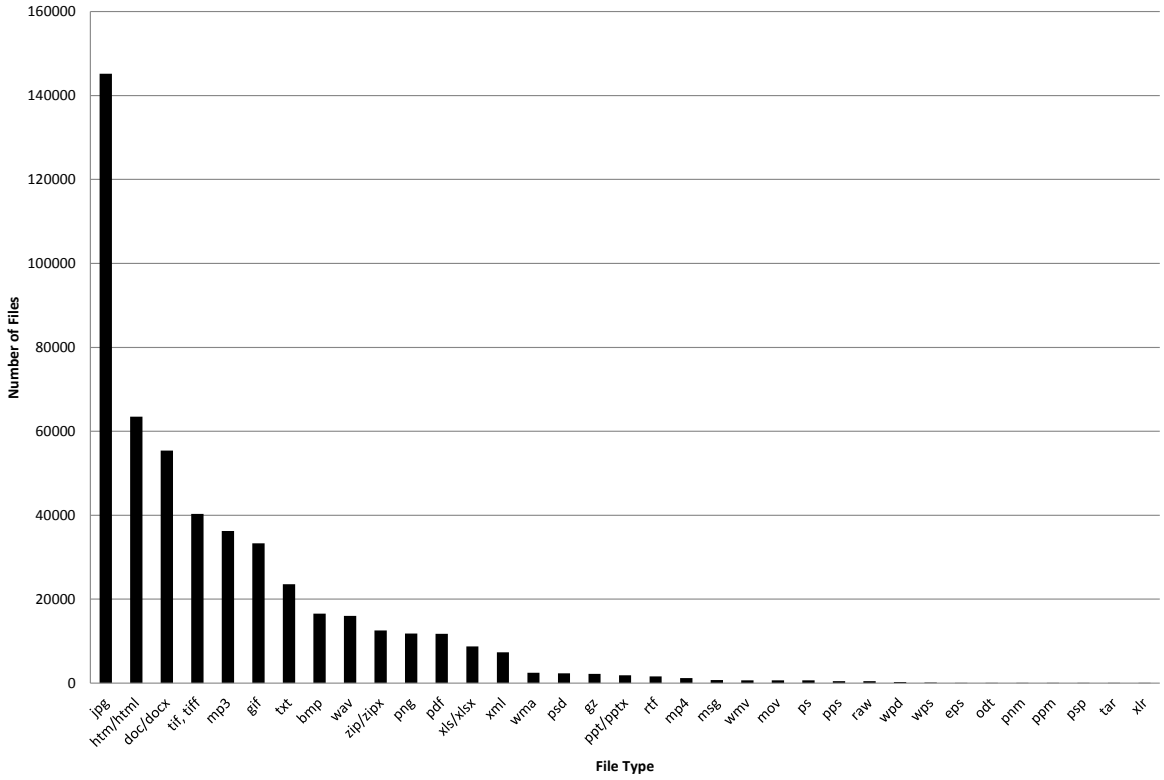


Figure 7.1: Corpus File Type Statistics.

ent information about user habits. These HTML files mainly consists of web pages saved to disk and user manuals for products. Since most of these documents are not generated by the user, they are not included. Each of these HTML files also likely includes an assortment of images. Browsing through the extracted GIF files reveals a number of simplistic graphics that are used for web page borders or have other decorative purposes. This high number of non-interest files demonstrates that, even with excluding common directories and pruning duplicates, there is still a certain level of noise.

This research extracted 205,389 images from the disks and found another 34,553 images embedded in Microsoft Word documents, as shown on Figure 7.2. To understand the applicability of this data set towards automated image annotation and latent topic extraction research, several tests were accomplished. Graphs, such as

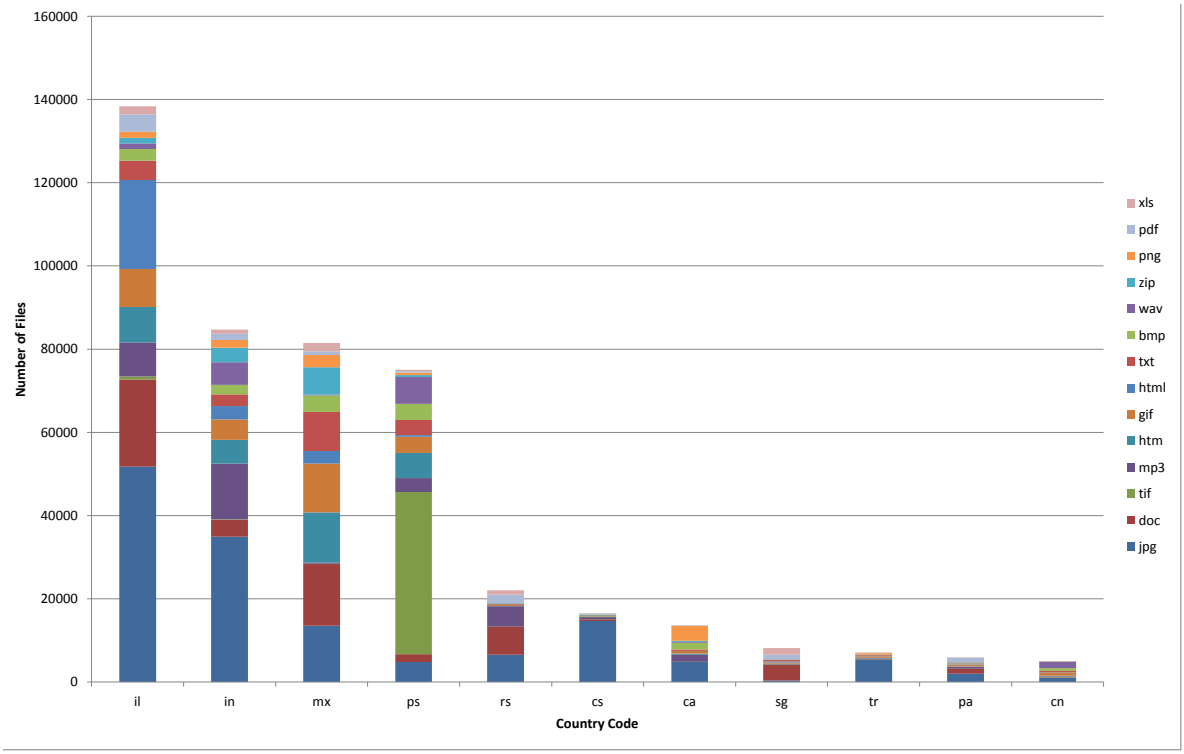


Figure 7.2: Corpus File Types by Country.

Table 7.2: Corpus Image Statistics for Regular and Embedded Images.

	Regular	Embedded
Image Count	205,389	34,553
Percent Graphs	35%	47%
Percent with faces	25%	11%
Percent with camera info	32%	N/A
Images with GPS Info	324	N/A

Table 7.3: Corpus Statistics for Images Embedded In Documents.

Number of documents	55,443
Perc. docs with embedded images	16%
Perc. docs with viable SWLDA images	8.5%
Disks with embedded image files	253
Avg embedded image file per disk	35

line graphs, pie charts, and line drawings often require different techniques than photographs since they lack complex image features. Each was checked to determine if it looked like a graph using the method defined in Section 4.3.2. The more shading or color variations in an image, the less likely it will be detected as a graph. Often, line drawings include shading or color shifts and result in a false negative. For this reason, the numbers listed in Table 7.2 are probably lower than the actual number.

The images extracted from the disk were examined for EXIF information. While many images had EXIF information indicating the editing software used to create or manipulate them, personal photographs tended to include camera information. Fully one-third of all images had camera information and a small percentage of those included Global Positioning System information from the location where the picture was taken.

This research considered attempting to automatically annotate images with local context, similar to Stage II. Unfortunately, as Table 7.3 indicates, only 16% of the Microsoft Word documents have embedded images with at most half those containing images that are not graphs. Considering that 253 disks have documents with embedded images, there were only an average of 35 viable documents per disk with no guarantee for cross-document topics. The low number of images results in a sparse super-word matrix and low probabilistic granularity. Due to the super-word influence and availability, automated image annotation using topical context is not attempted; however, this could provide an avenue for further image annotation research.

7.2.1.1 E-Mail and Web Caches.

The e-mail within the corpus was analyzed using topic models. Identifying e-mails from other text documents, such as logs, required using a regular expression search for e-mail addresses. An LDA model was built on the corpus of e-mails and analyzed. Unfortunately, due to the significant portion of advertisements and e-mail text from internet service providers, the latent topics did not reveal very interesting information.

The web caches were similarly analyzed with lackluster results. Most topics appeared to be clustering from advertisements or mail clients. While adult websites did cluster into their own topics, most other topics were difficult to draw many inferences from the results. Better results may be gained by utilizing filtering tools to eliminate advertising and boilerplate text before clustering with LDA.

The poor performance of e-mail and web analysis could be improved by using intelligent parsing algorithms. Web pages tend to include the main subject of their page in the center with advertisements, links, and artwork surrounding the topical text. E-mails tend to include specific formatting that could be used to separate manuals containing e-mail addresses from actual e-mails. Commercial tools exist that extract data from both e-mail and web cache data. Web tools provide browsing history and pattern information, as well as facilitate relevant data extraction from web pages. E-mail data extraction tools look for e-mails in common locations and formats. Future work may be improved by utilizing these pre-processing methods or commercial tools.

7.2.2 Using LDA for Digital Forensics.

The LDA algorithm was tested on the corpus using a variety of tasks mimicking standard digital investigations tasks. Our initial research focused on using LDA to

Table 7.4: Sample of LDA Results from India Corpus.

Topic	File Count	Representative Words
10	583	pune project ltd work com date training management pvt experience
11	249	india years bank company market year financial policy services
12	511	date prices power pune supply may required installation order
13	21	nanotubes carbon dna memory nram nanotube sequence computer
14	8	horus earth god seven two mother egyptian great heaven amenta
15	4	iii drone vii viii xii queen nic xvii xiii bel analogy ioc polaris patni xvi

extract common topics among various corpus. Table 7.4 illustrates the results of an LDA model built using the India corpus. Each topic is represented by a set of words that have the highest probability of being in that topic. Many topics, including 10-12 in the example, appear business-related. Topic 10 has a project management, personnel, or resumé tone, also mentioning the Indian city of Pune. Topic 11 contains words related to finance or market analysis. Topic 12 refers to the city of Pune, but appears to deal more with ordering systems. With topic 13 and 14, very different topics are evident from the rest–the first related to science and the second ancient Egyptian gods. The final topic listed only has four files associated with it and appears to include a number of Roman numerals.

Initial analysis of the Indian corpus seems to indicate the main topics center around business and science. The Israeli corpus includes some business topics as well as a large number of Jewish religious and cultural topics. It also contains several unique topics such as photography, educational topics, and Japanese historical and political analysis,

Table 7.5 defines an initial estimate of the major topics that emerged when LDA was performed on the corpus for each individual country. Most topics from the Mexican Word documents were in Spanish and the single English topic looked like it was derived from software manuals. Some, like the hard drives labeled from Panama, contained moderately uniform topics centered around medical research. A brief survey

Table 7.5: Country Microsoft Word Document Count and Major Topics.

Country (code)	Total Word Files	Assessed Topics
Israel (il)	210477	Jewish culture and religion, politics, news
Mexico (mx)	14777	all Spanish words, plus one computer topic
Serbia & Montenegro (rs/cs)	7565	commerce, power generation, health
India (in)	4332	business, science
Singapore (sg)	3833	business
Palestinian State (ps)	1921	computer systems/applications, business
Panama (pa)	1244	medical research
Pakistan (pk)	212	business
Ukraine (ua)	170	UAE business, French words
China (cn)	134	N/A
Unknown (nnn)	50	German words
Egypt (eg)	49	syringe production
Hong Kong (hk)	34	N/A
Turkey (tr)	34	N/A
Ghana (gh)	11	N/A
Canada (ca)	10	N/A
United Arab Emirates (ae)	7	N/A
Japan (jp)	3	N/A

of the document contents indicates significant amounts of Arabic letters, signifying these drives were likely mislabeled and came from Pakistan. The researchers who collected the data had awareness of this and had plans to correct the labels.

The inconsistent labels provide valuable data points for this research. Two LDA results for the disks labeled ‘pk’ and those labeled ‘pa’ offer some clues to how they were collected. While the ‘pk’ disks include large quantities of English words and business topics, the hard drives labeled ’pa’ contain numerous Arabic letters. Where there are English words, the topics are mainly focused on medical research and academics. Processing some of the Arabic text through a translator reveals medical topics. This implies that the drives marked ‘pa’ were drawn from similar businesses while the ones marked ‘pk’ were drawn from a medical university.

Country corpus with fewer than 200 documents tended to perform poorly, producing topics without distinct differences in topics. The drives marked ‘ua’ had a number

of documents related to business in the United Arab Emirates from four separate hard drives while a fifth hard drive contained mostly French words. The LDA model produced topics that had a mix of English and French words which, given those appeared in completely separate documents, illustrated a lack of consistent topical division in LDA. Likely, this was caused by the low number of documents and would benefit from some parameter tuning. The Chinese hard drive documents mostly contained Mandarin symbols which are out of scope for this research.

7.2.3 Test 1: Information retrieval.

The first test compares information retrieval results using regular expression search and LDA search. Documents are first pruned that have fewer than ten words and words that have fewer than ten instances in the corpus. Initial testing demonstrates this caused no decrease in performance for keyword search and is required for adequate performance of LDA search.

Since LDA is a stochastic algorithm, results will vary depending on the starting order of words and documents. To get an average, the LDA algorithm is run 30 times. Posterior probabilities provide the top 25% of documents most likely to match the keywords. These documents are used in a second iteration of the LDA algorithm where the process is performed again.

7.2.3.1 Retrieval topic 1: Passport files.

The first retrieval topic assumes the analyst wants to find documents related to passport requests. Four files have been pre-identified in the Israeli corpus and keywords selected (Table 7.6) that are similar to what an analyst might choose for this purpose. After pruning, 10,999 documents from 70 disks are able to be searched. Three of the target documents are in one disk while the fourth is in another.

Table 7.6: Topic 1 Keywords

Keywords	Weight
passport(s)	3
ambassador	3
embass(y|ies)	2
visa	2
clearance	1
application	1

Table 7.7: Topic 1 Document Rankings - Passport Files.

		Regex Search	LDA Iteration			
			1	2	3	4
Document #		10999	10999	2506	615	150
Doc 1	μ	11	1459.1	282.3	86.9	9.4
	σ	—	512.0	135.5	57.0	6.4
Doc 2	μ	58	1766.8	242.9	62.8	10.2
	σ	—	1139.8	168.0	51.0	6.3
Doc 3	μ	53	1859.5	297.1	68.1	10.2
	σ	—	1223.7	145.9	59.9	5.5
Doc 4	μ	N/A	2749.5	459.5	91.3	32.7
	σ	—	1515.0	231.5	50.3	30.2
Prob. Of Loss		—	0.49	0.37	0.12	0.44

Table 7.6 and 7.7 provide the list of keywords and the average position of the four target documents within the query results. Standard deviation provides the variation of the LDA algorithm results and may be an indicator of the difficulty the algorithm has categorizing the document. The final row, 'Prob. of Loss', provides the likelihood that a document will not survive to the next iteration, assuming only 25% of the documents survive.

Regular expression keyword search does better than a single application of the LDA query algorithm by returning the target documents in positions 11, 53, and 58 of the results. The fourth document is pruned from the list since it did not include any of the query words in the text. LDA, on the other hand, has a 0.50 probability of being pruned out on the first iteration. Assuming the fourth document survives to the second iteration, the smaller number of words in the topic decreases

the probability that the document would be pruned to 0.37. The fourth iteration of the LDA algorithm provides more relevant query results than regular expression search. Due to the small number of files remaining, the uncertainty in the fourth document re-emerges with a 0.44 probability of being pruned.

This experiment demonstrates LDA's advantage over keyword search. Keyword search does not identify the fourth document since none of the keywords were in it. Words topically similar to the keywords are in the document, however, so LDA has a moderate chance of returning it as a possible document.

7.2.3.2 Retrieval Topic 2: Legal Documents.

The second topic search attempts to retrieve documents for a court case involving a particular dispute between two families over the will of a deceased member. In this test, the name of the deceased is first included in the keyword search (Table 7.8), increasing the chance that a keyword search would be successful.

Table 7.8: Topic 2 Keywords - With Name.

Keywords	Weight
(masked name)	3
plantiff	2
defendent	2
estate	2
deceased	1
inheritance	1
marriage	1

Table 7.9 gives the results for the regular expression and LDA search using the whole Israel corpus. Keyword search returns relevant results, minus the first document. LDA is run and documents pruned four times but fails to provide more relevant results. Additionally, the second document has a moderate to high probability that it will be pruned off.

Table 7.10 demonstrates the same test, but removes the deceased's name from

Table 7.9: Topic 2 Document Rankings - Legal Documents With Name.

		Regex Search	LDA Iteration			
			1	2	3	4
Document #		10999	10999	2506	613	150
Doc 1	μ	153	578.0	235.9	64.9	29.2
	σ	—	325.2	152.9	24.8	12.8
Doc 2	μ	60	853.0	486.6	158.6	55.6
	σ	—	669.6	245.7	64.6	26.7
Doc 3	μ	1	951.8	146.7	49.2	16.4
	σ	—	802.8	195.9	29.5	12.0
Doc 4	μ	4	1409.1	122.2	30.3	14.1
	σ	—	1086.6	162.4	32.7	11.9
Prob. Of Loss		—	0.11	0.30	0.47	0.77

the search. In this test, keyword search fails to return the first document since it lacks any keywords while LDA returns it each iteration. LDA, unfortunately, has a moderately high probability of pruning the second document from the list.

Table 7.10: Topic 2 Keywords - Without Name.

Keywords	Weight
plantiff	2
defendent	2
estate	2
deceased	1
inheritance	1
marriage	1

The two sets of results in Tables 7.9 and 7.11 demonstrate a main advantages of LDA over keyword search. Since LDA is a topic model, documents are often returned even if there isn't a single keyword match. All that is required is for the keywords to be topically similar to words found within the document. One of the drawbacks to the LDA method over keyword search is that developing an effective LDA model typically requires word pruning. As the number of documents in the LDA model decreased, topically significant words were pruned. This meant the first two documents had a high probability of dropping off the list.

Since all four documents are from the same disk, Table 7.12 provides the keyword

Table 7.11: Topic 2 Document Rankings - Legal Documents Without Name.

		Regex Search	LDA Iteration			
			1	2	3	4
Document #		10999	10999	2481	614	152
Doc 1	μ	N/A	555.1	217.3	130.6	24.2
	σ	—	340.4	68.9	44.7	22.8
Doc 2	μ	523	999.0	662.8	222.5	31.6
	σ	—	571.4	240.7	92.7	15.9
Doc 3	μ	1	911.4	167.3	51.0	5.7
	σ	—	619.0	82.6	132.7	48.5
Doc 4	μ	7	1390.5	132.7	65.4	7.6
	σ	—	1072.1	48.5	25.4	5.0
Prob. Of Loss		—	0.11	0.57	0.78	0.63

Table 7.12: Topic 2 Document Rankings - Legal Documents Without Name, Single Disk.

		Regex Search	LDA Iteration		
			1	2	3
Document #		262	262	66	33
Doc 1	μ	20	5.8	18.1	22.7
	σ	—	2.9	15.1	6.7
Doc 2	μ	1	9.2	6.6	15.3
	σ	—	4.8	9.5	5.4
Doc 3	μ	10	3.0	1.5	1.4
	σ	—	2.3	0.5	0.5
Doc 4	μ	2	1.6	1.5	1.6
	σ	—	1.1	0.6	0.5
Prob. Of Loss		—	0.00	0.16	0.84

search and LDA results using just that disk. This results in far fewer documents than with the entire corpus and improves the keyword search results. It also requires fewer iterations of the LDA algorithm to return good LDA results. On the other hand, the user must search through twenty documents to find all relevant results and only ten, on average, using LDA. In later iterations, documents 1 and 2 seem to be losing some keywords that their ranking relies upon since they are descending the list. The final iteration was only pruned 50% since the number of documents was small. Without decreasing the prune rate, document 1 would have likely been pruned.

Table 7.13: Topic 2 Document Rankings - Legal Documents Without Name, 50% Retention.

		Regex	LDA Iteration								
		Search	1	2	3	4	5	6	7	8	9
Document #		10999	10999	5476	2745	1375	687	343	172	86	43
Doc 1	μ	N/A	383.3	423.4	213.8	150.6	79.9	68.2	41.6	30.4	19.9
	σ	—	245.0	151.5	76.6	77.4	43.8	31.7	24.1	13.2	1.0
Doc 2	μ	523	1532.2	720.3	324.0	171.7	117.1	58.3	37.2	18.3	15.7
	σ	—	654.1	222.5	163.3	81.8	90.6	28.2	12.2	7.7	3.0
Doc 3	μ	1	389.5	370.4	86.4	44.9	8.9	10.4	7.4	3.0	0.9
	σ	—	249.1	164.3	40.1	26.5	18.0	7.7	5.7	2.4	1.15
Doc 4	μ	7	253.9	258.2	101.3	70.2	16.4	10.3	9.2	4.2	1.0
	σ	—	119.0	106.3	20.5	37.3	16.9	6.9	8.7	5.9	0.8
Prob. Loss		—	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.17	0.14

LDA has trouble maintaining consistent topics for the four documents in question due to the large quantity of noisy data. However, as the corpus is pruned away, topics become cleaner and improve retrieval performance. This can be illustrated by using the second experiment on the Israel corpus without the name, but only pruning half of the documents each iteration. Table 7.13 has the results of all nine iterations, along with the likelihood of loss. As indicated, probability of loss is essentially zero for every step, demonstrating an improvement at the cost of processing time.

7.2.3.3 Retrieval Topic 3: Power Generation Documents.

Table 7.14: Topic 3 Keywords

Keywords	Weight
electricity	3
power	2
distribut(e|ion)	2
consumption	2
generat(or|ion)	2
network	1

The third topic search uses the corpus from Serbia (rs) and Montenegro (cs) to locate a set of seven technical documents describing electrical power distribution.

Table 7.15: Topic 3 Document Rankings - Power Distribution Documents.

		Regex Search	LDA Iteration			
			1	2	3	4
Document #		6,780	6,780	1,668	417	204
Doc 1	μ	70	139.4	98.0	98.8	148.3
	σ	—	73.3	38.8	66.2	19.5
Doc 2	μ	68	181.0	127.2	101.9	148.0
	σ	—	71.5	37.0	63.4	19.5
Doc 3	μ	560	101.7	114.0	95.2	108.5
	σ	—	41.4	37.8	45.2	27.5
Doc 4	μ	225	423.7	343.3	157.2	85.9
	σ	—	95.6	62.5	48.8	24.6
Doc 5	μ	61	394.3	346.8	154.5	85.4
	σ	—	64.4	61.3	48.3	24.9
Doc 6	μ	201	100.0	80.7	79.7	56.6
	σ	—	44.9	44.1	61.4	42.4
Doc 7	μ	179	374.7	281.1	191.2	121.7
	σ	—	74.5	45.9	48.6	30.3
Prob. Of Loss, 75% prune			0.00	0.13	0.86	1.00
Prob. Of Loss, 50% prune			0.00	0.00	0.15	0.99

Within two iterations, LDA places some of the documents higher in the query results than the regular expression. However, pruning to fewer than 200 documents risks pruning the fourth document. After the third iteration, the document prune rate is decreased from 75% to 50% to prevent pruning relevant documents. Ordinarily, predicting when to change the prune amount would be difficult. Pruning too many too early increases the chance of removing relevant documents. Pruning too few results in larger processing times or false positive rates. Unfortunately, pruning rate changes only help the algorithm survive one more. The next iteration has an almost guaranteed chance of losing the first two documents when pruning off 75% and very high probability when pruning 50%. Additionally, the document order swapped between iterations 3 and 4. Iteration 3 had low performance for documents 3 and 4 while iteration 4 had low performance for documents 1 and 2. This illustrates how important training can be to model performance.

Table 7.16: Topic 3 Document Rankings - Power Distribution Documents, Single Disk.

		Regex Search	LDA Iteration			
			1	2	3	4
Document #		6,075	6,075	1,513	387	194
Doc 1	μ	70	142.0	99.5	114.0	124.9
	σ	—	42.9	31.0	85.2	33.7
Doc 2	μ	68	177.2	127.7	116.9	126.8
	σ	—	41.4	24.4	83.2	32.8
Doc 3	μ	539	94.6	106.2	100.5	85.1
	σ	—	17.4	27.9	46.0	15.4
Doc 4	μ	224	357.5	328.1	139.5	122.8
	σ	—	60.9	79.2	44.8	24.5
Doc 5	μ	61	359.1	322.0	138.7	122.3
	σ	—	54.5	65.0	47.3	23.4
Doc 6	μ	200	76.7	78.4	88.8	68.1
	σ	—	26.3	42.6	53.4	20.6
Doc 7	μ	178	315.8	249.9	179.6	129.8
	σ	—	37.5	26.5	57.8	29.1
Prob. Of Loss, 75% prune			0.00	0.27	0.93	0.99
Prob. Of Loss, 50% prune			0.00	0.00	0.40	0.94

Table 7.16 lists the results of running regular expression search and LDA search on only the disk containing the target documents. This disk represents the majority of the Serbian and Montenegro corpus, containing 6,075 on disk versus 6,780 for the entire corpus. With the slightly smaller corpus, keyword search performs only marginally better while LDA search has slightly better results. The probability of pruning a relevant document is somewhat higher, mainly due to fewer documents.

The three information retrieval topics in this section demonstrate that LDA often provides improved query results, though word pruning should be managed to avoid pruning relevant documents. Keyword choice is important; however, LDA relaxes the requirement to have at least one keyword in the target document. As long as the words in the keyword list match the major topic for the document, it will likely be returned as a query result. Iterating LDA and pruning results improve document retrieval to a point, though there does appear to be a minimum number of documents required

to produce an effective model. While the threshold where this happens varies from model to model, it appears to be around 200 to 400 documents. For Tables 7.7 and 7.9, the results started to falter around 150 documents. Table 7.12 had decreased performance between 66 and 262 documents. Table 7.15 lost query performance around 500 documents. The optimal time to stop pruning will vary from model to model; however, finding a method to determine this point should improve search reliability.

7.2.4 Test 2: Subtopic Discovery.

Occasionally digital forensics practitioners may not have specific topics in mind, but want to understand the overall topics within a corpus. Traditional techniques manually view a sampling of the documents, or select a set of potential keywords to isolate expected topics. Delineating between topics, on the other hand, is a major strength of LDA. The India corpus has a large number of scientific documents from a variety of disciplines. The experiment described in this section tests the ability of LDA to isolate scientific documents amid other topics. Next, these documents are broken up into their natural latent sub-topics.

An LDA model is trained on the entire India corpus of Microsoft Word documents and non-Latin alphabet characters are discarded. Documents having fewer than ten words and words occurring fewer than ten times in the corpus are pruned. The words with the highest posterior probability for each topic are analyzed with an assessment made as to the overall topic. Four topics out of thirty included scientific words defining the topic. Documents that were most likely to belong to one of those four topics are selected and used for the second run of the LDA algorithm.

After a second LDA model is formed from the 828 potential scientific documents, an assessment is made on the subject of each topic and these are provided in Table

7.17. Each document is assessed to determine if it matches the anticipated subject. The 'Total' column lists the total number of documents assigned to that topic based on dominant probability. The 'Correct' column provides the number of documents that were determined to match the assessed subject. In some cases, two topics were determined to have the same subject and they are linked with equal values in the 'Group' column. 'Topic Percent' provides the number correct in the given topic while 'Category Percent' lists the number correct in the indicated group.

Table 7.17: LDA Science Topic Analysis of the Indian Corpus.

Topic	Estimated Topic	Group	Total	Correct	Topic Percent	Category Percent
0	Electronics	0	34	32	94.12%	94.12%
1	Quakes/Disaster Mgmt	1	6	4	66.67%	75.00%
17		1	6	5	83.33%	
2	Biology	2	7	6	85.71%	85.71%
4	Nanotubes	3	19	17	89.47%	89.47%
5	Materials/Stress	4	49	48	97.96%	99.35%
22		4	106	106	100.00%	
6	Navigation	5	9	7	77.78%	77.78%
7	Math	6	33	31	93.94%	93.94%
8	Chemistry	7	11	9	81.82%	93.55%
14		7	20	20	100.00%	
9	Medical	8	11	6	54.55%	64.29%
26		8	3	3	100.00%	
10	Vehicle Safety	9	18	17	94.44%	94.44%
11	Environmental	10	5	1	20.00%	78.57%
21		10	8	6	75.00%	
28		10	15	15	100.00%	
12	Electrical Power	11	82	82	100.00%	100.00%
18		11	227	227	100.00%	
13	Animal Science	12	14	3	21.43%	21.43%
15		13	0	0	0%	
16	Industrial Paints	14	5	5	100.00%	100.00%
29		14	78	78	100.00%	
19	Construction Science	15	13	13	100.00%	100.00%
23	Batteries	16	22	22	100.00%	100.00%
24	Color/Light	17	18	10	55.56%	55.56%
27	Engineering	18	8	8	100.00%	100.00%
				Total	84.30%	

Many of the documents were single pages scanned from an electronics textbook using optical character recognition. The LDA algorithm divided the chapter topics into broad electrical terms (topic 0), electrical chemistry (topic 14 and 23), electromagnetic induction (topic 12) and generators (topic 18). Topic 13 was thought to be 'Animal Science', but instead described lab procedures that happened to include handling of lab rabbits. Topic 24 was assessed as documents about color and light but included mostly job performance reviews. Some document topics were assessed and the results found to be similar, though with slight differences. For instance, topic 8 is assessed as chemistry, yet the documents were defining a chemical coating product.

The testing described in this section indicate that LDA can be used successfully to retrieve documents from a broad topic, then those topics can be divided into sub-categories. For an analyst conducting e-discovery, topics and sub-topics can be browsed similarly to a directory structure. Personal information could be ignored in favor of business or financial documents, and those documents divided and explored. In this instance, LDA correctly categorized documents over 80% of the time. Using the techniques defined in the first experiment, keywords can be used to retrieve specific documents from a desired sub-category.

7.2.5 Test 3: Overlapping Topic Analysis.

Words can have many meanings and the specific intent is often revealed by examining local context. Topics often overlap, with discussion of sub-topics such as disease relevant to both medicine and biological weapons. This overlap affects both search by keyword and LDA since overlapping topics could increase the false positive rates. Some topical complexities can be revealed by increasing the number of topics and causing a broad topic like "outdoor sports" to be categorized into "hunting" and "fishing" [46] [110]. A document about fishing, however, might be categorized as

both recreation and water conservation. LDA posterior probabilities may reveal both topics, but only if those topics have a large enough representation.

This test looks at topical shift as the number of topics is adjusted. As boundaries change, topical overlap can be identified. While simpler than [46] and [110], it offers a glimpse into the challenges of using topic models to browse document corpus. This used the "query by document" method where a document is submitted and LDA used to identify topically related documents. The LDA model was tested against a regular expression keyword search by first conducting a subjective analysis of the "known" document's topic. The keywords in Table 7.19 are derived from the pertinent information in the document. The document chosen for this test discusses government water policy.

The entire Indian corpus is used to estimate LDA model parameters, including the known document about water policy. Initially, 25 topics are used. The known document had 74% of its words assigned to a particular topic, so the 82 documents represented by that topic are extracted and used to form a new LDA model, similar to the methods used in test two. These documents were manually assigned labels by reviewing the documents and these are listed in Table 7.18. The labels in this topic are related to environmental management and agriculture, with a few unrelated topics in the 'Other' category. Still, a moderately large number of documents are unrelated to water.

Table 7.18 lists the topics derived from the LDA sub-model and reveals overlapping topics within the document. First, discussions of water includes governmental water policy, wastewater treatment, farming irrigation, water pollution, and using water in mining and mineral production. Government water policy is related to pollution since governmental regulations are often the remedy.

Table 7.20 offers some information about water topics and their topic shift as the

Table 7.18: Documents in the Water Topic.

Topic	Count
Water	15
Waste Management	15
Economy/Corporation/Trade	9
Environmental	8
Economy (farming)	6
Electricity	5
Agriculture	4
Transportation	3
Global Warming	3
Geography	2
Mining and Minerals	2
Seeds	2
Other	9

Table 7.19: Water Document Keywords.

Keywords	Weight
water	3
water resource(s)	3
united nations	2
international	1
management	1

number of topics is changed from 25 to 15 and 10. The document labels have been manually assigned with the same letter based on document similarity. Document 'A1' has related topics to the ones marked 'B' that primarily discuss water pollution. The 'E' and 'F' series documents are scientific documents discussing the removal of impurities from water, explaining the link with document 'A1' about pollution. Document 'D11' is a scientific analysis of irrigation techniques, explaining the link between that and the 'E' series with 15 topics, but also discusses governmental water policy as found in our 'C' series documents. Document 'C8' is the query document on governmental water policy.

Running a regular expression query using the keywords defined in Table 7.19 returns 73 documents and the results are provided in the last column of Table 7.20.

Table 7.20: LDA Clustering of Water Topic.

Document	Topics 25	15	10	Keyword Order
A1	7	1	6	22
B2	7	1	2	3
B3	7	1	2	20
B4	7	1	2	19
B5	7	1	2	21
B6	7	1	2	8
B7	7	1	2	6
C8	2	3	5	2
C9	2	3	5	1
C10	2	3	5	10
D11	2	9	14	7
E12	20	9	6	24
E13	20	9	6	30
F14	20	8	6	49
F15	20	8	6	23
X1	–	–	–	28

The regular expression query returns all documents provided by LDA, including one LDA did not find.

7.3 Conclusion

The techniques described in Stages I through III were tested on a data set that was not as noisy as data typically analyzed by digital forensics practitioners. In addition, the documents in a category were topically similar and provided a good data set to test topic models. This chapter analyzed real-world data from a digital forensics standpoint to determine which algorithms could prove useful to the analyst. First, statistics were gathered on the corpus to determine which algorithms may prove useful. Second, LDA search was compared against traditional regular expression search using three tests. The first test uses a set of queries on three different topics and demonstrated mixed results, with LDA performing better than regular expression

search if enough of the documents are pruned from the set. The second test examined topic and sub-topic browsing using LDA and demonstrated potential towards digital discovery. The third test demonstrated that LDA does provide slightly better query results than regular expression search in data with high topical overlap.

This chapter supports the fact that, while LDA is not a replacement for regular expression search, it offers some advantages. It relaxes the requirement to select keywords that match words within a target document. It also is able to effectively divide a corpus into topics and sub-topics and is resistant to topical overlap and noise.

VIII. Discussion and Conclusion

While data mining research has solved many problems in their attempts to extract useful information from data, many challenges still exist. The "large data problem" is a difficult problem, compounded by the lack of techniques utilizing multiple data domains. Images embedded in text documents offer additional context that can be used to develop representative models of the data. Many models, such as the generative Latent Dirichlet Allocation (LDA), provide the means for complex queries using model probability. Unfortunately, the model parameters are usually hidden and require estimation using the words and documents. Queries using the model will only be as effective as the accuracy of the model parameters in representing the data. Likewise, embedded images pose a similar challenge since they rarely come with accurate annotations, making queries for relevant images difficult. Chapter II provided much of the research focusing on improving models for the text or image domains separately.

This research hypothesizes that models combining text and embedded images can provide more relevant information than each domain individually. This hypothesis is broken up into three parts. First, this research addresses the hypothesis that automated image annotations can be improved using the local text surrounding an embedded image. Second, these automated image annotations can be used to improve latent topic models, providing better precision and recall in queries using unknown documents. Third, the posterior probabilities from the latent topic model can be used to further refine the image annotations.

Part I of this research presents novel methods for automatically annotating images and using these annotations to improve latent topic models. The model is described and tested using a three Stage model. In Stage I, context surrounding an image is used to prompt an automated image annotation algorithm, demonstrating improved

results over Automated Linguistic Indexing of Pictures in Real Time (ALIPR), a popular generalized image annotation algorithm. Stage II uses the image annotations from Stage I to train an LDA-based topic model, called Super-Word Latent Dirichlet Allocation (SWLDA), demonstrating improved precision and recall over many similar latent topic models. Stage III used the posterior probability from the SWLDA model to improve automated image annotation results over Stage I.

The data used for Part I was real-world data taken from a set of Wikipedia articles and provided a rich, yet topically clean, data set to test the three stages. Real-world user data from a digital forensics perspective will not be as clean or topically distinct. Part II of this research tests whether the models described in Part I can be used by the digital forensics analyst to extract meaningful information. To answer this question, several tests are performed that attempt to replicate common digital forensics practitioner tasks. This chapter first summarizes the three stages of Part I, then discusses the findings in Part II. It concludes with a discussion of future work.

8.1 Part I, Stage I: Context-Based Image Retrieval

Statistical pattern recognition assumes an overarching model is generating the data and its parameters can be estimated. In reality, the model representing real world data may be too complex for parameter extraction and data prediction. Test data that is manually selected risks introducing bias into the results of any experiment and image annotation research is no exception.

This research presented a method for annotating images, taking advantage of local image context to drive specialized image annotation algorithms using image features. By using an ensemble image annotation method to separate graphs, people-based images, and all other images, specialized annotators may be applied for each domain. The broad category of images that do not contain people and are not graphs is

annotated using a signature-based histogram comparison with Earth Mover's Distance (EMD). Each signature is generated using ImageNet images from a particular synset assisted by surrounding text. This algorithm was tested on a diverse assortment of real world data taken from the INitiative for the Evaluation of XML Retrieval (INEX) 2007 data set. Using this context-driven method outperforms ALIPR, a popular image annotation tool within the field.

While the findings of this research were promising, several questions remain. First, any useful system needs an acceptably low false positive rate to avoid flooding the user with spurious results. Currently there is no way of accurately determining how to draw out words that apply to the image and which do not. While this research made an attempt to experiment on data representing the chaotic nature of real world data, it still is not as unstructured and unpredictable as real user data corpus. Further research is required to determine how well this method maps to more complex data, but initial indications from Part II indicate it may be too noisy a domain, or that embedded images are the exception rather than the rule.

8.2 Part I, Stage II: Super Word LDA

Assuming the super-words have higher relevance, providing them greater influence can tighten categorical clustering and result in a statistically significant increase in recall at only a moderate loss of precision. Stage II demonstrated the SWLDA generative model for a document corpus consisting of low influence, high frequency words and high influence, low frequency super words. Unlike other multimodal LDA algorithms, this model leverages a common vocabulary to take advantage of the probabilistic coupling between words and super-words. This also helps with data smoothing, since super-words that do not have a word representation can be discarded.

The model was tested using super-words derived from the automated image anno-

tation technique defined in Stage I. Additionally, over 6,000 documents were drawn from the INEX 07 dataset [28] and clustered according to their specified categories. Comparison testing with Latent Semantic Analysis (LSA), LDA, Blob Multimodal Latent Dirichlet Allocation (BMMLDA), and Weighted Term Latent Dirichlet Allocation (WTLDA) showed that SWLDA offset a slight drop in precision with a higher recall and F-measure, especially as the number of topics was increased. The calculated precision, recall, and F-measure indicate an overall improvement using the SWLDA process over standard Gibbs sampling LDA.

While this dissertation tested the algorithm using super-words derived from images, it will work for any low density, high relevance words. This could include embedded sound with captions, table data, graph information, or other embedded data. While accurate super-words are important, this research has demonstrated it still can perform even with moderate super-word inaccuracy.

SWLDA was tested on documents containing text and images. The model could be used with audio converted into text and used to co-cluster with documents or a number of other media types. Section headers or captions can be used as super-words to improve clustering of documents. It could even be extended outside of the text domain, such as with applying LDA to large graphs [55]. If some of the nodes can be identified to be more relevant towards the applicable groupings than others, SWLDA can improve latent group clustering.

Once a model's parameters are predicted, super-word posterior probability can be used to rank-order automated image annotation. Those annotations without topical relevance are dropped while highly probable annotations are augmented with expanded synonyms. Iterating this process may further improve latent topic extraction. Initial testing has demonstrated promise.

8.3 Part I, Stage III: Image Annotation Refinement

Stage III demonstrated the posterior probabilities from Stage II could be used effectively to improve automated image annotations from Stage I. However, expanding the words and iterating the model further produced questionable results. These results require further Amazon Mechanical Turk (AMT) validation to determine if continued iteration is productive.

8.4 Part II: Topic Models in High Noise Data

Part II attempts to address whether the models described in this dissertation can be used by the digital forensics analyst to extract meaningful information from a corpus of real-world digital forensics data. First statistics were calculated on the user-generated data in the Real Data Corpus (RDC) to identify areas suitable for search evaluation. Using this information, a series of tests were conducted that replicate common digital forensics tasks. The first test used a set of queries on three different topics, comparing search using LDA to traditional regular expression search. The second test measured the ability of LDA to topically segment a country corpus, then separate a selected topic into subtopics. Finally, the third test compared query using LDA to query using regular expression in a topically noisy environment.

The RDC statistical analysis offers a variety of research targets. While many disks contain only operating system and application files, a number include a variety of documents, images, and sound files, among others. Many documents have embedded images and most images on disk include EXIF information. This, along with e-mails, log files, Excel spreadsheets and Powerpoint presentations, provide a rich data source for research. Documents are written in a variety of languages, some on the same disk. Most disks contain documents in the native language, but also include a number in English. With the focus of this research on comparing keyword search against LDA

search, this research constrained the data to English word or text documents.

LDA search demonstrated pros and cons over regular expression search. First, regular expression search often produced similar results within the first iteration, though LDA took four iterations. While pruning using LDA often resulted in higher precision due to smaller query results, it sometimes produced false negatives by pruning too much. Additionally, using LDA to search takes much longer. A regular expression search on the Israeli corpus took approximately one minute, yet four iterations of the LDA algorithm took over eight hours

On the other hand, LDA located documents even though the keywords used did not match any words within the document itself. LDA can find similar documents using a query document approach, unlike regular expression search which requires the query document be first converted into a keyword list. Additionally, LDA can automatically categorize documents into topics, permitting intelligent browsing of the document corpus.

8.5 Future Work

There are a number of potential avenues for further research. First, the corpus itself provides data useful in language processing, syntactical analysis, image processing, sound file analysis, and cross-domain techniques, among many others. Due to the manpower required to develop recall information within the RDC, this research limits its analysis to precision. Manually categorizing the documents within the corpus determines how many false negatives are produced by document query algorithms. Additionally, some tests required subjective analysis about document topics and could be improved by using a survey of varying opinions on document topics. Amazon Mechanical Turk has been successfully employed for this purpose in our previous research [90]. LDA and regular expression keyword search had strengths

that could be combined effectively into a hybrid technique.

During analysis of the results from Stage I and III, a number of erroneous annotations are passed on as valid and expanded. One example is the term 'wing' for an image of an aircraft. One definition that is commonly accepted by the algorithm is as "The fender of a motor vehicle". This is expanded into such words as 'bumper' and 'flap', which do not apply to the image of an aircraft. Ideally, the topic of a document about aircraft would not usually fit the use of these words to annotate the image. Future research could look into ways of further pruning these erroneous results, such as using WordNet clustering combined with topical posterior probability to eliminate poorly-fitting words.

This dissertation describes approaches that demonstrate that text surrounding an image can be used to improve automated image annotation. These annotations, in turn, can be used to improve latent topic extraction. While effective, data consists of many different formats beyond text and images. Future research could use text to improve sound file analysis, video analysis, and a myriad of different formats.

Besides different formats, specialized analyzers can be developed to analyze browsing or e-mail patterns. Attempts at analyzing web pages to determine browsing habits encountered significant noise. Most web pages include advertisements, links to related sites, and administrative links. Only a small portion of the page may actually be relevant to the user's purposes. These sections, however, may have common characteristics that can be leveraged to extract the useful information. For instance, most relevant user data may be in the center of the page and incorporate the majority of words. Images within this text may be relevant, whereas images along the edges of the web page may be advertisements that can be ignored. By pruning irrelevant data, SWLDA could be performed on the images using surrounding text to help annotate web images.

Regular expression search can be incorporated into the topic model processes to improve results. It can be used as a preprocessing step to prune unrelated results, or be used after a model has been generated on specific topics. Regular expression results could be combined with the posterior probability from the topic models to create a hybrid ranking of query results.

This research demonstrates leveraging cross-domain solutions can extract better information from a human perspective. Automated image annotation was improved by using the local context surrounding an embedded image. The annotations can be fed into the SWLDA algorithm to improve the recall and overall F-measure of the latent topic models. Finally, the posterior probability will improve the image annotation results.

Appendix A. Gibbs Sampling Super-Word Derivation

This derivation follows a similar approach to [16], building a proportional probability model to facilitate parameter estimation using Gibbs sampling and employs the assumption that words and super-words are drawn from a common vocabulary. Dirichlet priors β and η are assumed over the support of the entire corpus vocabulary. The probability distribution can be represented by Equation 1.1.

$$p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) = \prod_{k=1}^{K} p(\phi_k; \beta) p(\tau_k; \eta) \prod_{m=1}^{M} p(\theta_m; \alpha) \left(\prod_{n=1}^{N_m} p(z_{m,n}|\theta_m) p(w_n|\phi_k) \times \prod_{i=1}^{S_m} p(z_{m,N_m+i}|\theta_m) p(s_i|\tau_k) \right) \tag{1.1}$$

Assume the following distributions per Section V.

$$p(w_i|z_i, \phi^{(z_i)}) \sim Discrete(\phi^{(z_i)}), \tag{1.2a}$$

$$p(\phi) \sim Dirichlet(\beta), \tag{1.2b}$$

$$p(s_i|z_i, \tau^{(z_i)}) \sim Discrete(\tau^{(z_i)}), \tag{1.2c}$$

$$p(\tau) \sim Dirichlet(\eta), \tag{1.2d}$$

$$p(z_i|\theta^{(d_i)}) \sim Discrete(\theta^{(d_i)}), \tag{1.2e}$$

$$p(\theta) \sim Dirichlet(\alpha). \tag{1.2f}$$

The proportional equation with respect to z_i that facilitates sampling using a Monte-Carlo simulation appears in Equation 1.3. For the purposes of this derivation, the Bayesian denominator is not shown and normalization terms are calculated at the end.

$$p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) = \int\int\int p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) d\theta d\phi d\tau \tag{1.3}$$

Since θ and ϕ are isolated to their own terms, they can be separated into their own integrals.

$$\begin{aligned} p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) = & \prod_{m=1}^{M} \int p(\theta_m; \alpha) \prod_{n=1}^{N_m+S_m} p(z_{m,n}|\theta_m) d\theta_m \\ & \times \prod_{k=1}^{K} \int p(\phi_k; \beta) \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(z_{m,n}|\theta_m) p(w|\phi) d\phi_k \\ & \times \prod_{k=1}^{K} \int p(\tau_k; \eta) \prod_{m=1}^{M} \prod_{i=1}^{S_m} p(z_{m,i}|\theta_m) p(s|\tau) d\tau_k \end{aligned} \tag{1.4}$$

Expanding the Dirichlet and discrete distributions based on their definitions from Equations 1.2a - 1.2f results in:

$$\begin{aligned} p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) = & \prod_{m=1}^{M} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1} \prod_{n=1}^{N_m+S_m} \theta_{m,z_{m,n}} d\theta_m \\ & \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{m,k}^{\beta_v - 1} \prod_{m=1}^{M} \prod_{n=1}^{N_m} \phi_{z_{m,n}, w_{m,n}} d\phi_k \\ & \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} \tau_{m,k}^{\eta_v - 1} \prod_{m=1}^{M} \prod_{i=1}^{S_m} \tau_{z_{m,N_m+i}, s_{m,i}} d\tau_k. \end{aligned} \tag{1.5}$$

The innermost products of each term can be simplified using an exponent of word counts, allowing the θ, ϕ and τ products to be combined into one. The count variable is $c_{k,d,t,w}$ where k indicates the topics, d the document, t the token, and w a boolean determining either word or super-word.

$$
\begin{aligned}
p(z,w,s,\theta,\phi,\tau;\alpha,\beta,\eta) = & \prod_{m=1}^{M} \int \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k + c_{k,m,\bullet,\bullet}-1} d\theta_m \\
& \times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v + c_{k,\bullet,v,\gamma_w}-1} d\phi_k \\
& \times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\prod_{v=1}^{V}\Gamma(\eta_v)} \prod_{v=1}^{V} \tau_{k,v}^{\eta_v + c_{k,\bullet,v,\gamma_s}-1} d\tau_k
\end{aligned}
\tag{1.6}
$$

Multiplying each term by two fractions, each an inverse of the other that results in a product of one, allows sliding the integral over to isolate a new Dirichlet density.

$$
\begin{aligned}
p(z,w,s,\theta,\phi,\tau;\alpha,\beta,\eta) = & \prod_{m=1}^{M} \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(c_{k,m,\bullet,\bullet}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,\bullet,\bullet}+\alpha_k)} \\
& \int \frac{\Gamma(\sum_{k=1}^{K} c_{k,m,\bullet,\bullet}+\alpha_k)}{\prod_{k=1}^{K}\Gamma(c_{k,m,\bullet,\bullet}+\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k + c_{k,m,\bullet,\bullet}-1} d\theta_m \\
& \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \frac{\prod_{v=1}^{V}\Gamma(c_{k,\bullet,v,\gamma_w}+\beta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w}+\beta_v)} \\
& \int \frac{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w}+\beta_v)}{\prod_{v=1}^{V}\Gamma(c_{k,\bullet,v,\gamma_w}+\beta_v)} \phi_{k,v}^{\beta_v + c_{k,\bullet,v,\gamma_w}-1} d\phi_k \\
& \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V}\eta_v)}{\prod_{v=1}^{V}\Gamma(\tau_v)} \frac{\prod_{v=1}^{V}\Gamma(c_{k,\bullet,v,\gamma_s}+\eta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s}+\eta_v)} \\
& \int \frac{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s}+\eta_v)}{\prod_{v=1}^{V}\Gamma(c_{k,\bullet,v,\gamma_s}+\eta_v)} \tau_{k,v}^{\tau_v + c_{k,\bullet,v,\gamma_s}-1} d\tau_k
\end{aligned}
\tag{1.7}
$$

Since each integral represents an entire support of a Dirichlet distribution, it equals one. This means each integral can be dropped from the equation. In addition, since we're only seeking to maintain proportionality, the Dirichlet constant fractions at the beginning of each term can be dropped, greatly simplifying the equation (not possible if your Dirichlet priors are variable).

$$p(z,w,s,\theta,\phi,\tau;\alpha,\beta,\eta) \propto \prod_{m=1}^{M} \frac{\prod_{k=1}^{K}\Gamma(c_{k,m,\bullet,\bullet}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,\bullet,\bullet}+\alpha_k)} \times \prod_{k=1}^{K} \frac{\prod_{v=1}^{V}\Gamma(c_{k,\bullet,v,\gamma_w}+\beta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w}+\beta_v)} \times \prod_{k=1}^{K} \frac{\prod_{v=1}^{V}\Gamma(c_{k,\bullet,v,\gamma_s}+\eta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s}+\eta_v)} \tag{1.8}$$

Sampling draws a token $t_{d,e}$ from the document m. The goal is to isolate the terms dependent upon sample position (d,e), though that depends on whether $t_{d,e} \in W$ or $t_{d,e} \in S$. This requires the conditional formula in Equation 1.9.

$$p(z_{d,e}|z_{-(d,e)},w,s,\alpha,\beta,\eta) \propto \begin{cases} t_{d,e} \in W\text{:} & \prod_{m\neq d} \frac{\prod_{k=1}^{K}\Gamma(c_{k,m,\bullet,\bullet}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,\bullet,\bullet}+\alpha_k)} \times \frac{\prod_{k=1}^{K}\Gamma(c_{k,d,\bullet,\bullet}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,d,\bullet,\bullet}+\alpha_k)} \\ & \times \prod_{k=1}^{K} \frac{\prod_{v\neq w_{d,e}}\Gamma(c_{k,\bullet,v,\gamma_w}+\beta_v) \times \Gamma(c_{k,\bullet,w_{d,e},\gamma_w}+\beta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w}+\beta_v)} \\ & \times \prod_{k=1}^{K} \frac{\prod_{v\neq w_{d,e}}\Gamma(c_{k,\bullet,v,\gamma_s}+\eta_v) \times \Gamma(c_{k,\bullet,w_{d,e},\gamma_s}+\eta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s}+\eta_v)} \\ t_{d,e} \in S\text{:} & \prod_{m\neq d} \frac{\prod_{k=1}^{K}\Gamma(c_{k,m,\bullet,\bullet}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,m,\bullet,\bullet}+\alpha_k)} \times \frac{\prod_{k=1}^{K}\Gamma(c_{k,d,\bullet,\bullet}+\alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,d,\bullet,\bullet}+\alpha_k)} \\ & \times \prod_{k=1}^{K} \frac{\prod_{v\neq s_{d,e}}\Gamma(c_{k,\bullet,v,\gamma_w}+\beta_v) \times \Gamma(c_{k,\bullet,s_{d,e},\gamma_w}+\beta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w}+\beta_v)} \\ & \times \prod_{k=1}^{K} \frac{\prod_{v\neq s_{d,e}}\Gamma(c_{k,\bullet,v,\gamma_s}+\eta_v) \times \Gamma(c_{k,\bullet,s_{d,e},\gamma_s}+\eta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s}+\eta_v)} \end{cases} \tag{1.9}$$

By isolating a proportion based specifically on the selected d and e values along with the membership of $t_{d,e}$, terms (constants) can be dropped that don't depend upon d and e specifically. This results in the simplified Equation 1.10.

$$\propto \begin{cases} t_{d,e} \in W: & \frac{\prod_{k=1}^{K} \Gamma(c_{k,d,\bullet,\bullet} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,d,\bullet,\bullet} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\Gamma(c_{k,\bullet,w_{d,e},\gamma_w} + \beta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w} + \beta_v)} \times \\ & \prod_{k=1}^{K} \frac{\Gamma(c_{k,\bullet,w_{d,e},\gamma_s} + \eta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s} + \eta_v)} \\ t_{d,e} \in S: & \frac{\prod_{k=1}^{K} \Gamma(c_{k,d,\bullet,\bullet} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{k,d,\bullet,\bullet} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\Gamma(c_{k,\bullet,s_{d,e},\gamma_w} + \beta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w} + \beta_v)} \times \\ & \prod_{k=1}^{K} \frac{\Gamma(c_{k,\bullet,s_{d,e},\gamma_s} + \eta_v)}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s} + \eta_v)} \end{cases} \tag{1.10}$$

Next, the word indicated by d, e and f is removed from the equation by deriving an equivalent form. Based on the fact that $c = c^{-(d,e,f)} + 1$, we can add $+1$ to each place the word is removed to maintain equivalency between Equation 1.10 and 1.11 [16]. Terms that are not losing a word have $+1-1$ added to the term for eliminating the Γ term in a later step.

$$\propto \begin{cases} t_{d,e} \in W: & \frac{\prod_{k \neq z_{d,e}} \Gamma(c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_w)} + \alpha_k) \times \Gamma(c_{z_{d,e},d,\bullet,\bullet}^{-(d,e,\gamma_w)} + \alpha_{z_{d,e}} + 1)}{\Gamma(1 + \sum_{k=1}^{K} c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_w)} + \alpha_k)} \\ & \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,w_{d,e},\gamma_w}^{-(d,e,\gamma_w)} + \beta_{w_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w} + \beta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,w_{d,e},\gamma_w}^{-(d,e,\gamma_w)} + \beta_{w_{d,e}} + 1)}{\Gamma(1 + \sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_w}^{-(d,e,\gamma_w)} + \beta_v)} \\ & \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s} + \eta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}} + 1 - 1)}{\Gamma(1 - 1 + \sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_s} + \eta_v)} \\ t_{d,e} \in S: & \frac{\prod_{k \neq z_{d,e}} \Gamma(c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_s)} + \alpha_k) \times \Gamma(c_{z_{d,e},d,\bullet,\bullet}^{-(d,e,\gamma_s)} + \alpha_{z_{d,e}} + 1)}{\Gamma(1 + \sum_{k=1}^{K} c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_s)} + \alpha_k)} \\ & \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w} + \beta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}} + 1 - 1)}{\Gamma(1 - 1 + \sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_w} + \beta_v)} \\ & \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,s_{d,e},\gamma_s}^{-(d,e,\gamma_s)} + \eta_{s_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s} + \eta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,s_{d,e},\gamma_s}^{-(d,e,\gamma_s)} + \eta_{s_{d,e}} + 1)}{\Gamma(1 + \sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_s}^{-(d,e,\gamma_s)} + \eta_v)} \end{cases} \tag{1.11}$$

Based on the equivalency $\Gamma(x+1) = x \times \Gamma(x)$ [16], separated terms can be modified and Γ terms re-combined back into the product where possible. This results in Equation 1.12.

$$
\propto \begin{cases}
t_{d,e} \in W: & \frac{\prod_{k=1}^{K} \Gamma(c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_w)} + \alpha_k) \times (c_{z_{d,e},d,\bullet,\bullet}^{-(d,e,\gamma_w)} + \alpha_{z_{d,e}})}{\Gamma(1 + \sum_{k=1}^{K} c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_w)} + \alpha_k)} \\
& \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,w_{d,e},\gamma_w}^{-(d,e,\gamma_w)} + \beta_{w_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w} + \beta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,w_{d,e},\gamma_w}^{-(d,e,\gamma_w)} + \beta_{w_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_w}^{-(d,e,\gamma_w)} + \beta_v)} \\
& \times \frac{(c_{z_{d,e},\bullet,w_{d,e},\gamma_w}^{-(d,e,\gamma_w)} + \beta_{w_{d,e}})}{\sum_{v=1}^{V} (c_{z_{d,e},\bullet,v,\gamma_w}^{-(d,e,\gamma_w)} + \beta_v)} \\
& \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s} + \eta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}} - 1)}{\Gamma(-1 + \sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_s} + \eta_v)} \\
& \times \frac{(c_{z_{d,e},\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}} - 1)}{\sum_{v=1}^{V} (c_{z_{d,e},\bullet,v,\gamma_s} + \eta_v) - 1} \\
t_{d,e} \in S: & \frac{\prod_{k=1}^{K} \Gamma(c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_s)} + \alpha_k) \times (c_{z_{d,e},d,\bullet,\bullet}^{-(d,e,\gamma_s)} + \alpha_{z_{d,e}})}{\Gamma(1 + \sum_{k=1}^{K} c_{k,d,\bullet,\bullet}^{-(d,e,\gamma_s)} + \alpha_k)} \\
& \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_w} + \beta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}} - 1)}{\Gamma(-1 + \sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_w} + \beta_v)} \\
& \times \frac{(c_{z_{d,e},\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}} - 1)}{\sum_{v=1}^{V} (c_{z_{d,e},\bullet,v,\gamma_w} + \beta_v) - 1} \\
& \times \prod_{k \neq z_{d,e}} \frac{\Gamma(c_{k,\bullet,s_{d,e},\gamma_s}^{-(d,e,\gamma_s)} + \eta_{s_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{k,\bullet,v,\gamma_s} + \eta_v)} \times \frac{\Gamma(c_{z_{d,e},\bullet,s_{d,e},\gamma_s}^{-(d,e,\gamma_s)} + \eta_{s_{d,e}})}{\Gamma(\sum_{v=1}^{V} c_{z_{d,e},\bullet,v,\gamma_s}^{-(d,e,\gamma_s)} + \eta_v)} \\
& \times \frac{(c_{z_{d,e},\bullet,s_{d,e},\gamma_s}^{-(d,e,\gamma_s)} + \eta_{s_{d,e}})}{\sum_{v=1}^{V} (c_{z_{d,e},\bullet,v,\gamma_s}^{-(d,e,\gamma_s)} + \eta_v)}
\end{cases}
\tag{1.12}
$$

All the Γ terms can be considered constant, so are dropped. Remaining denominators are reduced to shorthand. Since β and η are constant symmetric priors, $\sum_{v=1}^{V} \beta_v$ is replaced with $V * \beta$. This results in the final derivation:

$$p(z, w, s, \theta, \phi, \tau; \alpha, \beta, \eta) \propto \begin{cases} t_{d,e} \in W: & \frac{(c^{-(d,e,\gamma_w)}_{z_{d,e},d,\bullet,\bullet} + \alpha_{z_{d,e}}) \times (c^{-(d,e,\gamma_w)}_{z_{d,e},\bullet,w_{d,e},\gamma_w} + \beta_{w_{d,e}})}{(c^{-(d,e,\gamma_w)}_{z_{d,e},\bullet,\bullet,\gamma_w} + V\beta)} \\ & \times \frac{(c_{z_{d,e},\bullet,w_{d,e},\gamma_s} + \eta_{w_{d,e}} - 1)}{(c_{z_{d,e},\bullet,\bullet,\gamma_s} + V\eta) - 1} \\ t_{d,e} \in S: & \frac{(c^{-(d,e,\gamma_s)}_{z_{d,e},d,\bullet,\bullet} + \alpha_{z_{d,e}}) \times (c^{-(d,e,\gamma_s)}_{z_{d,e},\bullet,s_{d,e},\gamma_s} + \eta_{s_{d,e}})}{(c^{-(d,e,\gamma_s)}_{z_{d,e},\bullet,\bullet,\gamma_s} + V\eta)} \\ & \times \frac{(c_{z_{d,e},\bullet,s_{d,e},\gamma_w} + \beta_{s_{d,e}} - 1)}{(c_{z_{d,e},\bullet,\bullet,\gamma_w} + V\beta) - 1}, \end{cases} \tag{1.13}$$

normalized over the sum of all K.

Appendix B. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) is a generative model that explains a set of observations using an unobserved set of probability distributions and assumed latent topics. It is an evolution from the Probabilistic Latent Semantic Analysis (pLSA) model [57] which can be considered an LDA model with multinomial priors generated using a uniform distribution. Since Blei, et al. defined the model [14] in 2003, it has been used and expanded in a number of different ways [48] [78] [109]. This appendix provides an intuitive explanation of the LDA model and associated concepts. It first discusses some of the terminology and concepts, then builds these into the overall LDA model.

B.1 Latent Topics and Latent Topic Models

A latent topic is a topic that is assumed present but is not currently visible. It typically defines topics within a text document corpus, but could be used to define topics within any information corpus. A topic is what a document is about conceptually, where the document associated with the topic would include words pertaining to the medical field or dog training.

Latent topic models assume these hidden topics exist within a collection of documents in different proportions. A single document may contain 30% of one topic and 70% of another, while a particular word may be generated by one topic 20% of the time and another 80%. These topics, incorporated into a probability model, provide predictive capabilities for unknown documents. By analyzing the words within the unknown document, the topics contained within the document can be predicted using a latent topic model.

For many latent topic models, a topic's subject is revealed using the words most likely to be generated by that topic. Table B.1 provides an example of an eleven-topic

Table B.1: LDA Topics with Highest Probability Associated Words.

Topic 1	ship	navy	war	class	hms	ships	royal
Topic 2	aircraft	wing	maximum	first	service	air	range
Topic 3	tank	gun	tanks	vehicle	war	used	also
Topic 4	first	one	also	dog	time	two	elephant
Topic 5	patrol	submarine	war	ship	two	navy	june
Topic 6	dog	dogs	breed	terrier	can	breeds	also
Topic 7	mountain	mount	peak	mountains	north	range	south
Topic 8	aircraft	flight	wing	air	can	also	used
Topic 9	plant	species	lightgreen	also	used	can	flower
Topic 10	building	tower	beinn	tallest	center	city	world
Topic 11	aircraft	air	missile	system	radar	range	force

LDA model that was generated from the same data used in Stages I through III of this dissertation. Based on the words that have the highest probability of belonging to that topic, we can infer certain subjects. Topics 1 and 5 are likely related to naval ships and warfare, while topics 2, 8 and 11 appear to discuss aircraft.

B.2 Generative Models and Processes

The LDA model is considered a generative model. A generative model assumes that the observable data was generated by a set of hidden statistical distributions. In the case of LDA, the model would generate documents and topics, then those documents and topics would generate words according to certain distributions. In reality, the generative process is purely theoretical since the observed data is typically all that is available. The model does, however, provide greater flexibility since it offers the ability to generate new data and determine if unknown data was generated by the model.

The LDA theoretical generative process is illustrated in B.1, where the number of words N is generated according to a Poisson distribution. The multinomial parameters for document topics, given as θ, are generated using a Dirichlet distribution with priors α. For each word that is generated, a topic z_n is selected based on the

multinomial distribution θ. The multinomial distribution ϕ_{z_n} for the word probability is dependent, or conditioned, on the topic that was selected.

Algorithm B.1 LDA Generative Process

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \phi)$, a multinomial probability conditioned on the topic z_n

B.3 Bayes Networks and the LDA Plate Diagram

Generative models using a Bayesian network are able to draw complex inferences using a network of random variables. Figure B.1 provides a simple Bayesian network that models the probability of a power brown-out. Power consumption depends, in part, on air conditioner usage and whether people are at home or work. The probability of it being hot outside depends partly on whether it is summer and daytime. Figure B.1 provides the Conditional Probability Table (CPT) for each random variable, indicating the likelihood that each event takes place. On any given day, there's a 5/7 chance of it being a weekday, or a 0.71 probability. There's a 0.80 probability of the weather being hot during a summer day while only a 0.01 probability during a non-summer night.

We can therefore calculate certain events, such as the probability of a brown-out given that it is a summer weekday

$$P(g|s, w) = \alpha \sum_h \sum_d P(g|h, w)P(h|s, d)P(d) \tag{2.14}$$

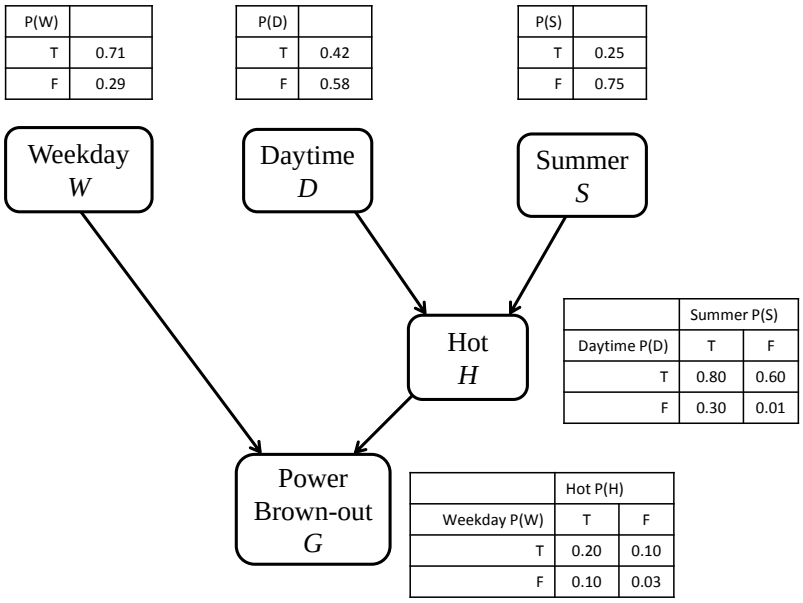


Figure B.1: Example Bayesian Network.

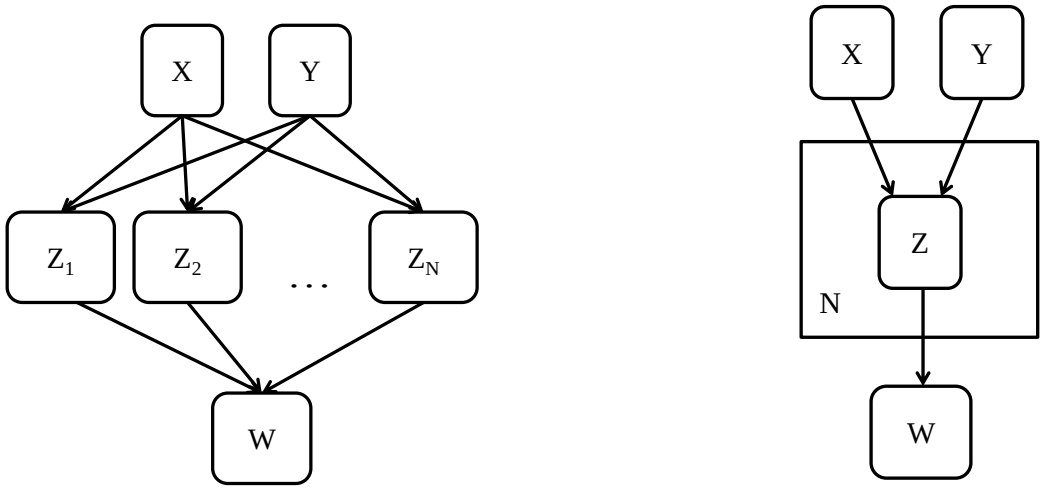
where α is a normalizing constant derived from the overall joint distribution. This results in a 0.15 probability that we are having a brown-out on a summer weekday. We can also determine the probability that it is hot outside, given that we are currently having a brown-out, or

$$p(h|g) = \frac{p(g|h)p(h)}{p(g)} \tag{2.15}$$

using Bayes rule. This results in

$$p(h|g) = \frac{\sum_w p(g|w,h)p(w)}{\sum_w \sum_h p(g|w,h)}. \tag{2.16}$$

Bayes networks can be described using a plate diagram. Similar to the example in B.1, a plate diagram includes nodes representing the elements of uncertainty and arrows representing dependencies. However, a plate diagram also indicates the number of times a graph segment repeats. Figures 2.2(a) and 2.2(b) illustrate the advantages of a plate diagram. Figure 2.2(a) illustrates a Bayesian network where z appears N times. If N was a large number, the graph becomes excessively complex.



(a) Bayes Network.

(b) Plate Diagram Version.

Figure B.2: Image Frequency Example.

Figure 2.2(b) illustrates the same Bayesian network, but does so in a much simpler fashion.

Figure B.3 illustrates the plate diagram for LDA. The outside plate indicates there are M instances of the multinomial distribution θ, or one per document. The θ distribution provides the probability that a topic z will be generated as topic 1 - K. There are N words, as indicated by the inner plate and each word has an associated topic z. The words are generated by K multinomial distributions, represented by ϕ . Finally, each multinomial distribution is generated using a Dirichlet distribution with priors α and β.

B.4 Latent Dirichlet Allocation Parameter Estimation

LDA is typically used to model a document corpus, where each document is considered a "bag of words" in which order does not matter. By using word co-occurrence, the model is able to identify topically similar documents and words by finding words that occur frequently in a subset of documents but not elsewhere. For instance,

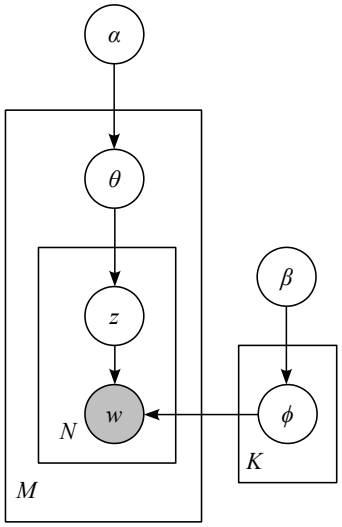


Figure B.3: Latent Dirichlet Allocation [14].

stethoscope, prescription, and diagnosis may occur frequently in medical documents, but will rarely occur in documents about music.

Note that solving for CPTs, in particular the hidden latent models, is not straightforward, since only the words are observable. Directly calculating the parameters is an NP-complete problem and requires using approximation techniques. One such technique, called Gibbs Sampling, uses a Monte-Carlo simulation to gradually adjust model parameters [48]. The process is outlined in Algorithm B.2.

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\bullet)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha} \tag{2.17}$$

Equation 2.17 is derived by reducing the joint probability distribution down to only those that are affected by changes in topic z_i with respect to word w_i. The result is a proportional sampling equation that can be used in the Gibbs sampler. The super-word sampling equation proof is defined in Appendix A and is similar to the proof for Equation 2.17.

Algorithm B.2 LDA Gibbs Sampling Algorithm

Input: words
Output: model parameters
Randomize document order and word order
Randomly assign topics to words
while not converged **do**
  for i=1 to N **do**
    Randomly select a topic based on existing parameters and Equation and assign to z_i for word w_i
    Adjust parameters
  end for
end while

B.5 LDA Inference

Once the model parameters are calculated, Bayes rule can be used to draw inferences about the data. The probability of a word given a topic, or $p(w|z)$ can be used to rank order the words associated with a particular topic and understand what that topic is about. The topic distribution of unknown documents can be calculated using Gibbs sampling or other techniques [54] to determine document topic probabilities.

Bibliography

[1] "OpenCV Computer Vision Library". C++ Library, May 2011.

[2] "U.S. Patent Office Data Dump", 2011. URL `http://www.uspto.gov`.

[3] Agarwal, Shivani, Aatif Awan, and Dan Roth. "Learning to Detect Objects in Images via a Sparse, Part-Based Representation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1475–1490, 2004.

[4] Aggarwal, Charu C. and ChengXiang Zhai. *Mining Text Data*. Springer, 2012.

[5] Andrews, Nicholas O. and Edward A. Fox. *Recent developments in document clustering*. Technical Report TR-07-35, Virginia Tech, 2007.

[6] Andrzejewski, David and David Buttler. "Latent Topic Feedback for Information Retrieval". *Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data Mining*, 600–608. ACM, 2011.

[7] Barnard, Kobus, Pinar Duygula, David Forsyth, Nando de Freitas, David M. Blei, and Michael I Jordan. "Matching Words and Pictures". *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[8] Barnard, Kobus, Pinar Duygulu, and David Forsyth. "Modeling the statistics of image features and associated text". *Conference on Document Recognition and Retrieval*, volume 4670, 1–11. 2002.

[9] Beebe, Nicole and Glenn Dietrich. "A new process model for text string searches". *Advances in Digital Forensics III*, 179–191. Springer, 2007.

[10] Beebe, Nicole Lang, Jan Guynes Clark, Glenn B. Dietrich, Myung S. Ko, and Daijin Ko. "Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies". *Decision Support Systems*, 51:732–744, 2011.

[11] Beecks, Christian, Merih Seran Uysal, and Thomas Seidl. "Signature Quadratic Form Distance". *ACM International Conference on Image and Video Retrieval*, 438–445. 2010.

[12] Blei, David and J. McAuliffe. "Supervised Topic Models". *Conference on Neural Information Processing Systems*, 1–8. MIT Press, 2007.

[13] Blei, David M. and Michael I. Jordan. "Modeling Annotated Data". *Special Interest Group on Information Retrieval*, 127–134. ACM, 2003.

[14] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3:993–1022, 2003.

[15] Busin, Laurent, Nicolas Vandenbroucke, and Ludovic Macaire. "Color spaces and image segmentation". *Advances in Imaging and Electron Physicxs*, 141:65–168, 2008.

[16] Carpenter, Bob. *Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling*. Technical Report, LingPipe, Inc, 2010.

[17] Carson, Chad, Serge Belongie, Hayit Greenspan, and Jitendra Malik. "Blobworld: Image segmentation using Expectation-Maximization and its application to image querying". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1026–1038, 2002.

[18] Chawla, Nitesh, Steven Eschrich, and Lawrence O. Hall. "Creating Ensembles of Classifiers". *First IEEE International Conference on Data Mining*, 580–581. 2000.

[19] Chen, Yixen and James Z. Wang. "A Region-Based Fuzzy Feature Matching Approach to Content-Based Image Retrieval". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1252–1267, 2002.

[20] Chen, Yixen, James Z. Wang, and Robert Krovetz. "CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning". *IEEE Transactions on Image Processing*, 14:1187–1201, 2005.

[21] Chia, Alex Y.S., Deepu Rajan, Maylor K.H. Leung, and Susanto Rahardja. "Category-Level Detection Based on Object Structures". *16th European Signal Processing Conference*, 1407–1421. 2008.

[22] Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. "Image Retrieval: Ideas, Influences, and Trends of the New Age". *ACM Computing Surveys*, 40(2):1–60, April 2008.

[23] Daubechies, Ingrid. "Ten Lectures on Wavelets". *Society for Industrial and Applied Mathematics*, 284–289. 1992.

[24] Davey, Monica. "Prosecutors Offer Blagojevich Evidence", April 2010. URL `http://www.nytimes.com/2010/04/15/us/politics/15blagojevich.html`.

[25] Dawson, J.L. "Suffix Removal for Word Conflation". *Bulletin of the Association for Literary and Linguistic Computing*, 2(3):33–46, 1974.

[26] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. "Indexing by Latent Semantic Analysis". *Journal of the American Society for Information Science*, 41:391–407, 1990.

[27] Deng, Jia, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". *IEEE Conference on Computer Vision and Pattern Recognition*, 785–792. 2009.

[28] Denoyer, Ludovic and Patrick Gallinari. "The Wikipedia XML Corpus". *Special Interest Group on Information Retrieval*, 12–19. 2007.

[29] Dietterich, Thomas G. "Ensemble methods in machine learning". *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, 97–136. 2000.

[30] Du, Lan, Huidong Jin, Olivier de Vel, and Nianjun Liu. "A latent semantic indexing and wordnet based information retrieval model for digital forensics". *IEEE International Conference on Intelligence and Security Informatics*, 70 – 75. IEEE, 2008.

[31] Dy, Jennifer G., Carla E. Brodley, Avi Kak, Chi-Ren Shyu, and Lynn S. Broderick. "The customized queries approach to CBIR using EM". *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 400–406. 1999.

[32] Farrahi, Katayoun and Daniel Gatica-Perez. "Learning and predicting multimodal daily life patterns from cell phones". *Proceedings of the 2009 International Conference on Multimodal Interfaces*, 277–280. ACM, 2009.

[33] Feldman, Susan and Chris Sherman. *The high cost of not finding information.* White paper, International Data Corporation, 2003.

[34] Fellbaum, Christiane. *WordNet: An electronic lexical database.* MIT Press, Cambridge, MA, 1998.

[35] Felzenszwalb, Pedro F. and Daniel P. Huttenlocher. "Efficient Graph-Based Image Segmentation". *International Journal of Computer Vision*, 59:167–181, 2004.

[36] Feng, Yansong and Mirella Lapata. "Automatic Image Annotation Using Auxiliary Text Information". *Proceedings of the Association for Computational Linguistics*, 951–955. 2008.

[37] Fleck, Margaret, David Forsyth, and Chris Bregler. "Finding Naked People". *European Conference on Computer Vision*, 593–602. Springer Berlin Heidelberg, 1996.

[38] Frawley, William J., Gregory Piatetsky-Shapiro, and Christopher J. Matheus. "Knowledge Discovery in Databases: An Overview". *AI Magazine*, 13:57–70, 1992.

[39] Furnas, G.W., T.K. Landauer, L.M. Gomez, and S.T. Dumais. "The vocabulary problem in Human-System communications". *Communications of the ACM*, 30:964–971, 1987.

[40] Garfinkel, Simson, Paul Farrell, Vassil Roussev, and George Dinolt. "Bringing science to digital forensics with standardized forensic corpora". *Digital Investigation*, 6:S2–S11, 2009.

[41] Garfinkel, Simson L. "Digital forensics research: The next 10 years". *Digital Investigations*, 7:64–73, 2010.

[42] Garfinkel, Simson L. "Real Data Corpus". Electronic, 2011. URL `http://digitalcorpora.org`.

[43] Gesù, Vito Di and Valery Starovoitov. "Distance-based functions for image comparison". *Pattern Recognition Letters*, 20:207–214, 1999.

[44] Giacinto, Giorgio. "A nearest-neighbor approach to relevance feedback in content-based image retrieval". *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, 456–463. 2007.

[45] Gonzalez, Rafael C., Richard E. Woods, and Steven L. Eddins. *Digital Image Processing Using MATLAB*. Prentice Hall, 2004.

[46] Gormley, Matthew R., Mark Dredze, Benjamin Van Durme, and Jason Eisner. "Shared components topic models". *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 783–792. Association for Computational LOinguistics, 2012.

[47] Grauman, Kristen and Trevor Darrell. "Efficient Image Matching with Distributions of Local Invariant Features". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 627–634. 2005.

[48] Griffiths, Thomas and Mark Steyvers. "Finding scientific topics". *Proceedings of the National Academy of Sciences*, 5228–5235. PNAS, 2004.

[49] Hansen, Lars Kai and Peter Salamon. "Neural Network Ensembles". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 1990.

[50] Haralick, Robert M., K. Shanmugam, and Its'hak Dinstein. "Textural Features for Image Classification". *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3:610–621, 1973.

[51] Harman, Donna. "How Effective is Suffixing". *Journal of the American Society for Information Science*, 42:1–7, 1991.

[52] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[53] Hauptmann, Alexander G. and Michael G. Christel. "Successful Approaches in the TREC Video Retrieval Evaluations". *Proceedings of the ACM Multimedia*, 668–675. 2004.

[54] Heinrich, Gregor. *Parameter estimation for text analysis*. Technical Report, Fraunhofer IGD, 2005.

[55] Henderson, Keith and Tina Eliassi-Rad. "Applying Latent Dirichlet Allocation to Group Discovery in Large Graphs". *Proceedings of the 2009 ACM Symposium on Applied Computing*, 1456–1461. ACM, 2009.

[56] Hofmann, Thomas. "The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data". *Proceedings of International Joint Conferences on Artificial Intelligence*, 682–687. 1999.

[57] Hofmann, Thomas. "Probabilistic Latent Semantic Indexing". *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, 50–57. ACM, 1999.

[58] Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining". *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20:19–62, 2005.

[59] Hull, D.A. "Stemming Algorithms - A Case Study for Detailed Evaluation". *Journal of the American Society for Information Science*, 47:1–70, 1996.

[60] Jeon, J., V. Lavrenko, and R. Manmatha. "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models". *Special Interest Group on Information Retrieval*, 119–126. 2003.

[61] Ji, Rongrong, Hongxun Yao, Zhen Zhang, Peifei Xu, and Jicheng Wang. "Using Visual Dictionary to Associate Semantic Objects in Region-Based Image Retrieval." *International Conference on Image Analysis and Recognition*, 615–625. 2007.

[62] Jiang, Wei, Guihua Er, Qionghai Dai, and Jinwei Gu. "Similarity-Based Online Feature Selection in Content-Based Image Retrieval". *IEEE Transactions on Image Processing*, 15:702–712, 2006.

[63] Jin, Yohan, Latifur Khan, Lei Wang, and Mamoun Awad. "Image annotations by combining multiple evidence & wordNet". *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, 706–715. ACM, New York, NY, USA, 2005. ISBN 1-59593-044-2.

[64] Jing, Yushi and Shumeet Baluja. "VisualRank: Applying PageRank to Large-Scale Image Search". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1877–1890, 2008.

[65] Jones, Karen Spärck. "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation*, 28:11–21, 1972.

[66] Jones, Michael and James M. Rehg. "Statistical color models with application to skin detection". *International Journal of Computer Vision*, 46:81–96, 2002.

[67] Kavitha, Ch, B. Prabhakara Rao, and A Govardhan. "An Efficient Content-Based Image Retrieval Using Color and Texture of Image Sub-Blocks". *International Journal of Engineering Science and Technology*, 3(2):1060–1068, February 2011.

[68] Kohonen, Teuvo. "The Self-Organizing Map". *Proceedings of the IEEE*, 78:1464–1480, 1990.

[69] Kutner, Michael H. *Applied Linear Statistical Models*. McGraw-Hill, 5 edition, 2005.

[70] Larlus, Diane, Jakob Verbeek, and Frédéric Jurie. "Category Level Object Segmentation by Combining Bag-of-Words Models with Dirichlet Processes and Random Fields". *International Journal of Computer Vision*, 88:238–253, 2009.

[71] Larson, Troy. "The other side of civil discovery". Eoghan Casey (editor), *Handbook of Computer Crime Investigation*, 17–52. Academic Press, 2002.

[72] Leong, Chee Wee, Rada Mihalcea, and Samer Hassan. "Text Mining for Automatic Image Tagging". *International Conference on Computational Linguistics*, 647–655. 2010.

[73] Li, Jia and James Z. Wang. "Real-Time Computerized Annotation of Pictures". *IEEE Transactions on Pattern Analysis and Machine Itelligence*, 30:985–1002, 2008.

[74] Li, Jia, James Z. Wang, and Gio Wiederhold. "IRM: Integrated Region Matching for Image Retrieval". *Proceedings of the 8th ACM International Conference on Multimedia*, 147–156. 2000.

[75] Li, Yi and Linda G. Shapiro. "Object Recognition for Content-Based Image Retrieval". *Lecture Notes in Computer Science*. Springer-Verlag, 2004.

[76] Lieberman, Henry and Hugo Liu. "Adaptive Linking Between Text and Photos Using Common Sense Reasoning". *In Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, 2–11. 2002.

[77] Liu, Hong, Yihua Lan, Xiangyang Xu, Enmin Song, and Chih-Cheng Hung. "Fissures Segmentation Using Surface Features: content-based retrieval for mammographic mass using ensemble classifier". *Academic Radiology*, 18:1475–1484, 2011.

[78] Liu, Jiakai, Rong Hu, Meihong Wang, Yi Wang, and Edward Y Chang. "Web-Scale Image Annotation". *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information*, 663–674. 2008.

[79] Liu, Ting, Charles Rosenberg, and Henry A. Rowley. "Clustering Billions of Images with Large Scale Nearest Neighbor Search". *IEEE Workshop on Applications of Computer Vision*, 28. IEEE, 2007.

[80] Liu, Ying, Dengsheng Zhang, Goujun Lu, and Wei-Ying Ma. "A survey of content-based image retrieval with high-level semantics". *Pattern Recognition*, 40(1):262–282, January 2007.

[81] Ma, Wei-Ying and B.S. Manjunath. "NeTra: A toolbox for navigating large image databases". *Multimedia Systems*, 7:184–198, 1999.

[82] Makadia, Ameesh, Vladimir Pavlovic, and Sanjiv Kumar. "A New Baseline for Image Annotation". *European Conference on Computer Vision (ECCV)*. 2008.

[83] McCullagh, Declan. "Finding treasures in Bin Laden computers", May 2011. URL `http://www.cbsnews.com/8301-501465_162-20060347-501465.html`.

[84] Mears, Bill. "Ex-Rep Rick Renzi of Arizona loses appeal on corruption charges", June 2011. URL `http://articles.cnn.com/2011-06-23/justice/arizona.renzi.appeal_1_corruption-charges-debate-clause-appeals-court?_s=PM:CRIME`.

[85] Mueller, Lance. "Crafting good keywords in EnCase and using conditions to refine results". ForensicKB, March 2013. URL `http://www.forensickb.com/2013/03/crafting-good-keywords-in-encase-and.html`.

[86] Müller, Henning, Stephane Marchand-Maillet, and Thierry Pun. "The truth about Corel – evaluation in image retrieval". *Proceedings of the Challenge of Image and Video Retrieval*, 38–49. 2002.

[87] Müller, Henning, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. "A review of content-based image retrieval systems in medical applications–clinical benefits and future directions". *International Journal of Medical Informatics*, 73:1–23, 2004.

[88] Nance, Kara, Brian Hay, and Matt Bishop. "Digital Forensics: Defining a Research Agenda". *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 1–6. IEEE, 2009.

[89] Napolitano, Janet. "The Future of Science as Public Service". Speech presented at the Massachusetts Institute of Technology, March 2011.

[90] Noel, George E. and Gilbert L. Peterson. "Context-Based Image Annotation Using ImageNet". *26th International Florida Artificial Intelligence Research Society Conference*, 462–467. AAAI, 2013.

[91] Opitz, David and Richard Maclin. "Popular Ensemble Methods: An Empirical Study". *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[92] Park, Bo Gun, Kyoung Mu Lee, and Sang Uk Lee. "Color-based image retrieval using perceptually modified hausdorff distance". *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[93] Pavlidis, Theo. "Why meaningful automatic tagging of images is very hard". *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, 1432–1435. 2009. ISBN 978-1-4244-4290-4.

[94] Perotte, Adler, Nicholas Bartlett, No'emie Elhadad, and Frank Wood. "Hierarchically Supervised Latent Dirichlet Allocation". *Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 12–15. MIT Press, 2011.

[95] Popescu, Adrian, Pierre-Alaim Moëllic, and Christophe Millet. "SemRetriev: An Ontology Driven Image Retrieval System". *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, 113–116. 2007.

[96] Quercia, Daniele, Harry Askham, and Jon Crowcroft. "TweetLDA: Supervised Topic Classification and Link Prediction in Twitter". *Proceedings of the 4th ACM International Conference on Web Science*, 247–250. ACM, 2012.

[97] Reisinger, Joseph, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. "Spherical Topic Models". *Proceedings of the 27th International Conference on Machine Learning*, 903–910. 2010.

[98] Reith, Mark, Clint Carr, and Gregg Gunsch. "An Examination of Digital Forensics Models". *International Journal of Digital Evidence*, 2002.

[99] Robertson, Philip K. "Visualizing Color Gamuts: A User Interface for the Effective Use of Perceptual Color Spaces in Data Displays". *IEEE Computer Graphics & Applications*, 8:50–64, 1988.

[100] Roth, Peter M. and Martin Winter. *Survey of Appearance-Based Methods for Object Recognition.* Technical Report ICG-TR-01/08, Graz University of Technology, Austria, 2008.

[101] Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. "The Earth Mover's Distance as a Metric for Image Retrieval". *International Journal of Computer Vision*, 40:99–121, 2000.

[102] Sakre, Mohammed M., Mohammed M. Kouta, and Ali M. N. Allam. "Weighting Query Terms Using WordNet Ontology". *International Journal of Computer Science and Network Security*, 9:349–358, 2009.

[103] Salton, G., A. Wong, and C.S. Yang. "A Vector Space Model for Automatic Indexing". *Information Retrieval and Language Processing*, 18:613–620, 1975.

[104] Schanda, J'anos. *Colormetry*. John Wiley & Sons, 2007.

[105] Schiele, Bernt, Mykhaylo Andriluka, Nikodem Majer, Stefan Roth, and Christian Wojek. "Visual People Detection - Different Models, Comparison and Discussion". *Proceedings of the IEEE International Conference on Robotics and Automation*, 1–8. 2009.

[106] Schober, Jean-Pierre, Thorsten Hermes, and Otthein Herzog. "Content-based image retrieval by ontology-based object recognition". *Workshop on Applications of Description Logics*, 61–67. 2004.

[107] Sebe, N., Q. Tian, E. Loupias, M. Lew, and T. Huang. "Evaluation of Salient Point Techniques". *Image and Vision Computing*, 21:1087–1095, 2003.

[108] Seemann, Edgar, Bastian Leibe, and Bernt Schiele. "Multi-Aspect Detection of Articulated Objects". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1582–1588. 2006.

[109] Shafiei, M. Mahdi. *Leveraging Structural Information for Statistical Topic Models of Text*. Ph.D. thesis, Dalhousie University, 2009.

[110] Shafiei, M. Mahdi and Evangelos E. Milios. "Latent Dirichlet Co-Clustering". *Proceedings of the Sixth International Conference on Data Mining*, 542–551. IEEE, 2006.

[111] Shi, Jianbo and Jitendra Malik. "Normalized Cuts and Image Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[112] Smeulders, Arnold W.M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. "Content-Based Image Retrieval at the End of the Early Years". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[113] Sridhar, Akshay, Scott Doyle, and Anant Madabhushi. "Boosted Spectral Embedding (BOSE): Applications to Content-Based Image Retrieval of Histopathology". *IEEE International Symposium on Biomedical Imaging*, 1897–1900. 2011.

[114] Steinbach, Michael, George Karypis, and Vipin Kumar. *A Comparison of Document Clustering Techniques*. Technical Report TR-00-034, University of Minnesota, 2000.

[115] Stricker, Markus and Michael Swain. "The Capacity of Color Histogram Indexing". *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 704–708. 1994.

[116] Swain, Michael J. and Dana H. Ballard. "Color Indexing". *International Journal of Computer Vision*, 7:11–32, 1991.

[117] Szummer, M. and R.W. Picard. "Indoor-Outdoor Image Classification". *IEEE International Workshop on Content-Based Access of Images and Video Databases*, 42–51. 1998.

[118] Tamura, Hideyuki, Shunji Mori, and Takashi Yamawaki. "Textural Features Corresponding to Visual Perception". *IEEE Transactions on Systems, Man and Cybernetics*, 8:460–473, 1978.

[119] Tang, Cheng-Yuan, Jeng-Horng Chang, and Chao-Tsung Hsieh. "Using color distribution and indexing for content-based image retrieval". *16th IPPR Conference on Computer Vision, Graphics and Image Processing*, 630–637. 2003.

[120] Vailaya, A., A. Jain, and H.J. Zhang. "On Image Classification: City Images vs. Landscapes". *Pattern Recognition*, 31:1921–1935, 1998.

[121] Velivelli, Atulya, Chong-Wah Ngo, and Thomas S. Huang. "Detection of documentary scene changes by audio-visual fusion". *Proceedings of the 2nd International Conference on Image and Video Retrieval*, 227–237. 03.

[122] Vezhnevets, Vladimir, Vassili Sazonov, and Alla Andreeva. "A Survey on Pixel-Based Skin Color Detection Techniques". *Proceedings on GRAPHICON*, 85–92. 2003.

[123] Viola, Paul and Michael J. Jones. "Rapid Object Detection using a Boosted Cascade of Simple Features". *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1511–1518. IEEE, 2001.

[124] Wallach, Hanna M., David Mimno, and Andrew McCallum. "Rethinking LDA: Why Priors Matter". *Proceedings of NIPS*, 1973–1981. 2009.

[125] Wang, Changhu, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. "Image Annotation Refinement using Random Walk with Restarts". *Proceedings of the 14th Annual ACM International Conference on Multimedia*, 647–650. ACM, 2006.

[126] Wang, Changhu, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. "Content-Based Image Annotation Refinement". *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8. 2007.

[127] Wang, Gang, Derek Hoiem, and David Forsyth. "Building text features for object image classifications". *Conference on Computer Vision and Pattern Recognition*, 1367–1374. IEEE, 2009.

[128] Wang, Huan, Song Liu, and Liang-Tien Chia. "Does ontology help in image retrieval?: A comparison between keyword, text ontology and multi-modality ontology approaches". *Proceedings of the 14th annual ACM International Conference on Multimedia*, 109–112. 2006. ISBN 1-59593-447-2.

[129] Wang, James, Gio Wiederhold, and Oscar Firschein. "System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms". *Computer Communications Journal*, 21:1355–1360, 1998.

[130] Wang, James Z., Jia Li, and Gio Wiederhold. "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:947–963, 2001.

[131] Wang, Xuerui, Andrew McCallum, and Xing Wei. "Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval". *Proceedings of the Seventh IEEE International Conference on Data Mining*, 697–702. 2007.

[132] Wangming, Xu, Wu Jin, Liu Xinhai, Zhu Lei, and Shi Gang. "Application of Image SIFT features to the context of CBIR". *International Conference on Computer Science and Software Engineering*, 552–555. 2008.

[133] Wei, Xing and W. Bruce Croft. "LDA-Based Document Models for Ad-hoc Retrieval". *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178–185. ACM, 2006.

[134] Westerveld, Thijs. "Image Retrieval: Content versus Context". *6th Conference on Content-Based Multimedia Information Access*, 276–284. 2000.

[135] Wilson, Andrew T. and Peter A. Chew. "Term weighting schemes for Latent Dirichlet Allocation". *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 465–473. 2010. ISBN 1-932432-65-5.

[136] Yang, Yi and Shawn Newsam. "Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery". *IEEE International Conference on Image Processing*, 1852–1855. 2008.

[137] Yoo, Illhoi and Xiaohua Hu. "A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE". *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 220–229. 2006.

[138] Zeger, Scott L. and M. Rezaul Karim. "Generalized Linear Models with Random Effects; A Gibbs Sampling Approach". *Journal of the American Statistical Association*, 86(413):79–86, 1991.

[139] Zhai, Yun, Alper Yilmaz, and Mubarak Shah. "Story segmentation in news videos using visual and textual cues". *Proceedings of the ACM Multimedia*, 3568:92–102, 2005.

[140] Zhang, Qi, Sally A. Goldman, Wei Yu, and Jason E. Fritts. "Content-Based Image Retrieval Using Multiple-Instance Learning". *International Workship on Machine Learning*, 682–689. 2002.

Glossary

Bayesian network A probabilistic graphic model that represents random variables and conditional dependencies using a graph. The network facilitates inferences about certain events given other events have occurred.

Daubechies wavelets A family of orthogonal wavelets that provide a localized frequency measures using a fast wavelet transform.

Dirichlet distribution A distribution based off the beta distribution that represents the probability of k rival events.

Earth Mover's Distance A measure designed to compare two histograms for similarity. It measures the amount of "earth" that must be moved to turn one histogram into another.

ensemble director An element of a model that chooses classifiers based on attributes of the data to be classified.

F-measure A harmonic mean between precision and recall that can be weighted to prefer one over the other.

generative model A model that assumes observable data was generated from a typically hidden model.

hypernym A more general form of another word. This word has a "is-a" relationship with another word. Example: 'mammal' is a hypernym of 'dog'.

hyponym A more specific form of another word. This word has a "type-of" relationship with another word. Example: 'dog' is a hyponym of 'mammal' as it is a more general form.

image blobs Spatially-contiguous regions of an image that have similar color or frequency attributes. Examples include regions of sky, grass, or a horse.

ImageNet An online database of images organized by WordNet synonymous sets. Includes features useful in data mining research, from bounding boxes to SIFT features.

Latent Dirichlet Allocation A generative model that explains a set of observations using multinomial distributions. The multinomial distributions can be explained using Dirichlet distributions.

latent topic model A model that assumes hidden topics exist within the data.

Markov chain monte-carlo A method to determine hidden model parameters using observations generated by the model. By randomly sampling data conditioned on previously-sampled data, the parameters can be estimated.

multimodal data Data consisting of multiple formats, such as image and text or video and audio.

multinomial distribution A probability distribution capable of representing n independent trials with k categories. Each has a fixed probability.

plate diagram A Bayesian network diagram that includes plates signifying instances of probabilistic events.

posterior probability The probability of an event assigned after an event or number of similar events have taken place.

precision Pertaining to document retrieval queries, precision indicates the percent of in-category documents retrieved out of all documents retrieved.

Real Data Corpus A set of user's data drawn from real hard drives acquired on the open market. The corpus was built by the Naval Postgraduate School and University of North Carolina.

recall Pertaining to document retrieval queries, recall indicates the percent of in-category documents retrieved out of all in-category documents within the corpus.

regular expression A sequence of characters, including metacharacters with special meaning, that define a search pattern.

summarization Reducing the amount of data or number of data dimensions into a representative subset.

super-word A word that is assumed to have higher contextual relevance than a regular word.

synset A WordNet term describing a synonymous set of words, or words that have similar meaning.

WordNet An online dictionary that organizes words into synonymous sets, or synsets. It includes hierarchical relationships between synsets using hypernyms and hyponyms, among others.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12/09/2013	Dissertation	20/08/2010 - 12/09/2013

4. TITLE AND SUBTITLE
Image Annotation and Topic Extraction Using Super-Word Latent Dirichlet Allocation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Noel, George E. LtCol

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
Wright-Patterson AFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENG-DS-13-S-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/RIGB
Chad Heitzenrater
525 Brooks Rd.
Rome Labs AFB, NY 13441
chad.heitzenrater@afr

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RIGB

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A. Approved For Public Release; Distribution Unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This research presents a multi-domain solution that uses text and images to iteratively improve automated information extraction. Stage I uses local text surrounding an embedded image to provide clues that help rank-order possible image annotations. These annotations are forwarded to Stage II, where the image annotations from Stage I are used as highly-relevant "super-words" to improve extraction of topics. The model probabilities from the super-words in Stage II are forwarded to Stage III where they are used to refine the automated image annotation developed in Stage I. All stages demonstrate improvement over existing equivalent algorithms in the literature.

15. SUBJECT TERMS
Latent Dirichlet Allocation, Topic models, Automated Image Annotation, Data Mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Gilbert L. Peterson, AFIT/ENG
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	181	(937) 255-3636 x4281 gilbert.peterson@afit.edu